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

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

Abstract

We present NeRSP, a Neural 3D reconstruction technique for Reflective surfaces with Sparse Polarized images. Reflective surface reconstruction is extremely challenging as specular reflections are view-dependent and thus violate the multiview consistency for multiview stereo. On the other hand, sparse image inputs, as a practical capture setting, commonly cause incomplete or distorted results due to the lack of correspondence matching. This paper jointly handles the challenges from sparse inputs and reflective surfaces by leveraging polarized images. We derive photometric and geometric cues from the polarimetric image formation model and multiview azimuth consistency, which jointly optimize the surface geometry modeled via implicit neural representation. Based on the experiments on our synthetic and real datasets, we achieve the state-of-the-art surface reconstruction results with only 6 *views as input.*

1. Introduction

Multiview 3D reconstruction is a fundamental problem in computer vision (CV) and has been extensively studied for many years [14]. With the advancement of implicit surface representation [27, 28] and neural radiance fileds [22], recent multiview 3D reconstruction methods [5, 33, 38, 41] have made tremendous progress. Despite the compelling shape recovery results, most multiview stereo (MVS) methods still rely heavily on finding correspondence between views, which is particularly challenging for reflective surfaces and sparse input views.

For reflective surfaces, the view-dependent surface appearance breaks the photometric consistency assumption used in the correspondence estimation in MVS. To address this problem, recent neural 3D reconstruction methods (*e.g.*, Ref-NeuS [13], NeRO [19], and PANDORA [9]) explicitly model the reflectance and simultaneously estimate the reflectance and environment maps via inverse rendering. However, dense image acquisition under diverse views is required to faithfully handle the additional unknowns besides shape, such as albedo, roughness, and environment map.

From sparse input views, it is often challenging to find

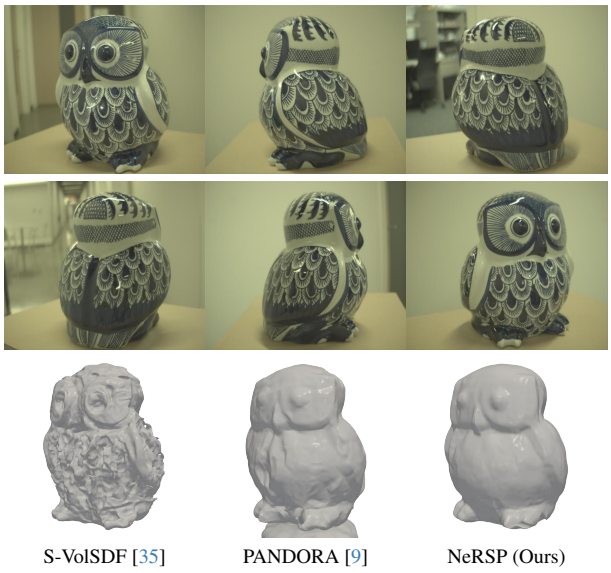

Figure 1. Shape recoveries of a reflective surface from 6 sparse polarized images capturing (top rows). Our NeRSP achieves a better shape reconstruction result compared to existing methods that either address sparse inputs (S-VolSDF [35]) or reflective reflectance (PANDORA [9]).

† Equal contribution. * Corresponding author.
Project page: https://yu-fei-han.github.io/NeRSP-project/

sufficient multiview correspondences. Especially when representing view-dependent reflectances, it is difficult to disentangle shape from radiance under a limited number of correspondences, leading to shape-radiance ambiguity [40]. Recent neural 3D reconstruction methods for sparse views (*e.g.*, S-VolSDF [35] and SparseNeuS [20]) require regularization using photometric consistency, which can be violated for reflective surfaces.

To address both problems, we propose to use sparse polarized images instead of RGB inputs. Specifically, we propose NeRSP, a Neural 3D reconstruction method to recover the shape of Reflective surfaces from Sparse Polarized images. We use the angle of polarization (AoP) derived from polarized images, which directly reflects the azimuth angle of the surface shape up to π and $\pi/2$ ambiguities. This geometric cue is known to enable multiview shape reconstruction regardless of surface reflectance properties, but the estimated shape based solely on the geometric cue is ambiguous [6] under sparse views settings. On the other hand, photometric cue from the polarimetric image formation model [2] helps neural surface reconstruction (*e.g.*, PANDORA [9]) by minimizing the difference between rerendered and captured polarized images. However, estimated shape based solely on the photometric cue is also ill-posed under sparse inputs due to the shape-radiance ambiguity. Unlike the existing polarimetric-based method PANDORA [9] considering the photometric cue only, our NeRSP shows the integration of both geometric and photometric cues effectively narrows down the solution space for surface shape, shown to be effective in reflective surface reconstruction based on sparse inputs, as visualized in Fig. 1.

Besides the proposed NeRSP for 3D reconstruction, we also build a Real-world MultiView Polarized image dataset containing 6 objects with aligned ground-truth (GT) 3D meshes, named RMVP3D. Different from existing datasets such as PANDORA dataset [9] providing polarized images only, the aligned GT meshes and the surface normals for each view allow a quantitative evaluation of multiview polarized 3D reconstruction.

To summarize, we advance multiview 3D reconstruction by proposing

- NeRSP, the first method proposing to use the polarimetric information for reflective surface reconstruction under sparse views;
- a comprehensive analysis for the photometric and geometric cue derived from polarized images; and
- RMVP3D, the first real-world multiview polarized image dataset with GT shapes for quantitative evaluation.

2. Related work

Multiview 3D reconstruction has been extensively studied for decades. Neural Radiance Fields (NeRF) [3, 22, 40] achieves great success on novel view synthesis in recent years. Inspired by NeRF, neural 3D reconstruction methods [24] are proposed, where the surface shape is modeled implicitly via signed distance field (SDF). Beginning from DVR [24], the followed-up methods improve the shape reconstruction quality via differentiable sphere tracing [37], volume rendering [26, 33, 38], or detail enhanced shape representation [18, 34]. These methods can achieve convincing shape estimation for diffuse surfaces where photometric consistency is valid across views.

3D reconstruction for reflective surfaces is challenging as the photometric consistency is invalid. Existing methods [5, 41, 42] explicitly model the view-dependent reflectance, and disentangle the shape, spatially-varying illuminations, reflectance properties like albedo and roughness. However, the estimates of the above variables are open unsatisfactory as the disentanglement is highly ill-posed. NeRO [19] proposes using the split-sum approximation of the image formation model and further improves shape reconstruction quality without requiring object masks. However, the above methods typically require dense image capture to guarantee plausible shape recovery results for challenging reflective surfaces.

3D reconstruction with sparse views is essential for practical scenarios requiring efficient capture. Due to the lack of sufficient correspondence from limited views, the shape-radiance ambiguity cannot be resolved, leading to noisy and distorted shape recoveries. Existing methods address this problem by adding regularizations such as surface geometry smoothness [25], coarse depth prior [10, 32], or frequency control of the positional encoding [36]. Some methods [7, 20, 39] formulate the sparse 3D reconstruction as a conditioned 3D generalization problem where image features pre-trained are used as generalizable priors. S-VolSDF [35] applies classical multiview stereo method as initialization and regularizes the neural rendering optimization with a probability volume. However, it is still challenging for current methods to recover reflective surfaces accurately.

3D reconstruction using polarized images has been studied for both single view setting [1, 2, 16, 23, 29] and multiview setting [6, 8, 9, 11, 12, 43]. Unlike RGB images, the AoP from polarized images provides direct cues for surface normal. Single-view shape from polarization (SfP) techniques benefit from this property and estimate the surface normal under single distant light [21, 29] or unknown natural light [1, 16]. Multiview SfP methods [8, 43] resolve the π and $\pi/2$ ambiguities in the AoP based on the multiview observations. PANDORA [9] is the first neural 3D reconstruction method based on polarized images, demonstrated to be effective in recovering surface shape and illumination. MVAS [6] recovers surface shape from multiview azimuth maps, closed related to the AoP maps derived from

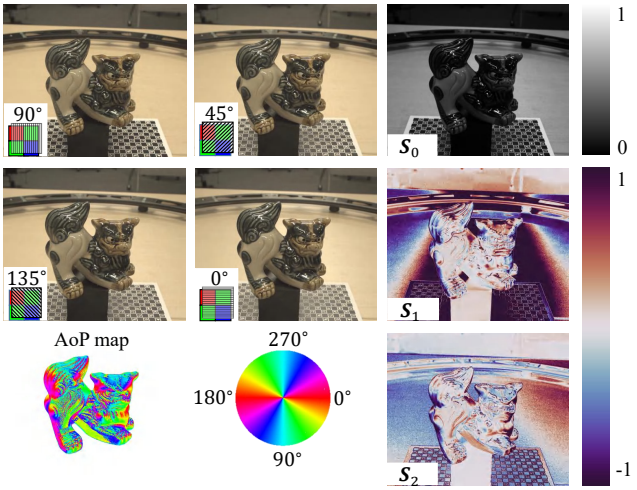

Figure 2. Visualization of polarized images, derived AoP map, and Stokes vectors.

polarized images. However, these methods do not explore using polarized images for reflective surface reconstruction under sparse shots.

3. Polarimetric Image Formation Model

Before dive into the proposed method, we first introduce polarimetric image formation model and derive the photometric cue and geometric cue in our method.

As shown in Fig. 2, a snapshot polarization camera records image observations at four different polarization angles, with its pixel values denoted as $\{I_0, I_{45}, I_{90}, I_{135}\}$. This four images reveal the polarization state of received lights, which is represented as a 4D Stokes vector $\mathbf{s} = [s_0, s_1, s_2, s_3]$ computed as

$$\begin{cases} s_0 = \frac{1}{2}(I_0 + I_{45} + I_{90} + I_{135}) \\ s_1 = I_0 - I_{90} \\ s_2 = I_{45} - I_{135}. \end{cases} \tag{1}$$

We assume there is no circularly polarized light thus assign s_3 to be 0. The Stokes vector can be used to compute the angle of polarization (AoP), *i.e.*

$$\phi_a = \frac{1}{2}\arctan(\frac{s_2}{s_1}). \tag{2}$$

Based on the AoP and Stokes vector, we derive the geometric and photometric cue correspondingly.

3.1. Geometric cue

Given AoP ϕ_a, the azimuth angle of the surface can be either $\phi_a + \pi/2$ or $\phi_a + \pi$, known as the π and $\pi/2$ ambiguity depending on whether the surface is specular or diffuse dominant. In this section, we first introduce the geometric cue brought by multiview azimuth map, and then extend it to the case of AoP.

Following MVAS [6], for a scene point $\mathbf{x}$, its surface normal $\mathbf{n}$ and the projected azimuth angle ϕ in one camera view follow the relationship as

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

where $\mathbf{R} = [\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3]^\top$ is the rotation matrix of the camera pose. We can further re-arrange Eq. (3) to get the orthogonal relationship between surface normal and a projected tangent vector $\mathbf{t}(\phi)$ as defined below,

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

The π ambiguity between AoP and azimuth angle 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 k 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 multiview polarized 3D reconstruction.

3.2. Photometric cue

Assuming the incident environment illumination is unpolarized, the Stokes vector of the incident light direction $\boldsymbol{\omega}$ can be represented as

$$\mathbf{s}_i(\boldsymbol{\omega}) = L(\boldsymbol{\omega})[1, 0, 0, 0]^\top, \tag{7}$$

where $L(\boldsymbol{\omega})$ denotes the light intensity. The outgoing light recorded by the polarization camera becomes partially polarized due to the reflection. This process is modeled via a 4×4 Muller matrix $\mathbf{H}$. Under an environment illumination, the outgoing Stokes vector $\mathbf{s}_o$ can be formulated as the integral of incident Stokes vector multiplicated with the Muller matrix, *i.e.*

$$\mathbf{s}_o(\mathbf{v}) = \int_\Omega \mathbf{H}\mathbf{s}_i(\boldsymbol{\omega})\ d\boldsymbol{\omega}, \tag{8}$$

where $\mathbf{v}$ and Ω denote the view direction and integral domain. Following the polarized BRDF (pBRDF) model [2], 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.*

$$\mathbf{s}_o(\mathbf{v}) = \int_\Omega \mathbf{H}_d\mathbf{s}_i(\boldsymbol{\omega})\ d\boldsymbol{\omega} + \int_\Omega \mathbf{H}_s\mathbf{s}_i(\boldsymbol{\omega})\ d\boldsymbol{\omega}. \tag{9}$$

Following the derivation from PANDORA [9], we can furhter formulate the output Stokes vector as

$$\mathbf{s}_o(\mathbf{v}) = L_d \begin{bmatrix} T_o^+ \\ T_o^- \cos(2\phi_n) \\ -T_o^- \sin(2\phi_n) \\ 0 \end{bmatrix} + L_s \begin{bmatrix} R^+ \\ R^- \cos(2\phi_h) \\ -R^- \sin(2\phi_h) \\ 0 \end{bmatrix}, \quad (10)$$

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 related to surface normal $\mathbf{n}$, Fresnel transmission coefficients [2] $T_{i,o}^+$ and $T_{i,o}^-$, diffuse albedo ρ, and the azimuth angle of incident light ϕ_n. $L_s = \int_\Omega L(\boldsymbol{\omega}) \frac{DG}{4\mathbf{n}^\top \mathbf{v}}\, d\boldsymbol{\omega}$ denotes specular radiance related to Fresnel reflection coefficients [2] R^+ and R^-, the incident azimuth angle ϕ_h w.r.t. the half vector $\mathbf{h} = \frac{\boldsymbol{\omega}+\mathbf{v}}{\|\boldsymbol{\omega}+\mathbf{v}\|_2^2}$, and the normal distribution and shadowing term D and G in the Microfacet model [31]. Please check supplementary material for more details.

Based on polarimetric image formation model shown in Eq. (10), we build the photometric cue.

4. Proposed method

Our NeRSP takes sparse multiview polarized images, the corresponding silhouette mask of the target object, and camera poses as input, outputs surface shape of the object represented implicitly via SDF. We begin by the discussion on photometric cue and geometric cue in resolving the shape reconstruction ambiguity, followed by the instruction of network structure and loss function of our NeRSP.

4.1. Ambiguity in sparse 3D reconstruction

The geometric cue and photometric cue play an important role in reducing the solution space of the surface shape under sparse views. As shown in Fig. 3, we illustrate the shape estimation under 2 views with different cues. Given only RGB images as input (corresponding to the setting in NeRO [19] and S-VolSDF [35]), different combination of scene point positions, surface normals, and reflectance properties such as albedo can lead to the same image observations, since there are only two RGB measurements for each 3D points along the camera ray. With Stokes vectors extracted from the polarized images, the photometric cue brings 6 measurements for each 3D points (Stokes vector has 3 elements), reducing the surface normal candidates unfit to the polarimetric image formation model.

On the other hand, based on AoP maps[1] from polarized images, we can uniquely determine the surface normal up to a π ambiguity for every scene point along the camera ray. However, it is still ambiguous to find the position where camera ray intersects the surface, unless the third view is provided [6]. Therefore, under sparse views setting (*e.g.*, 2 views in Fig. 3), determining scene point position based on either geometric or photometric cue remains ambiguous.

[1] AoP is related to the azimuth map discussed in MVAS [6].

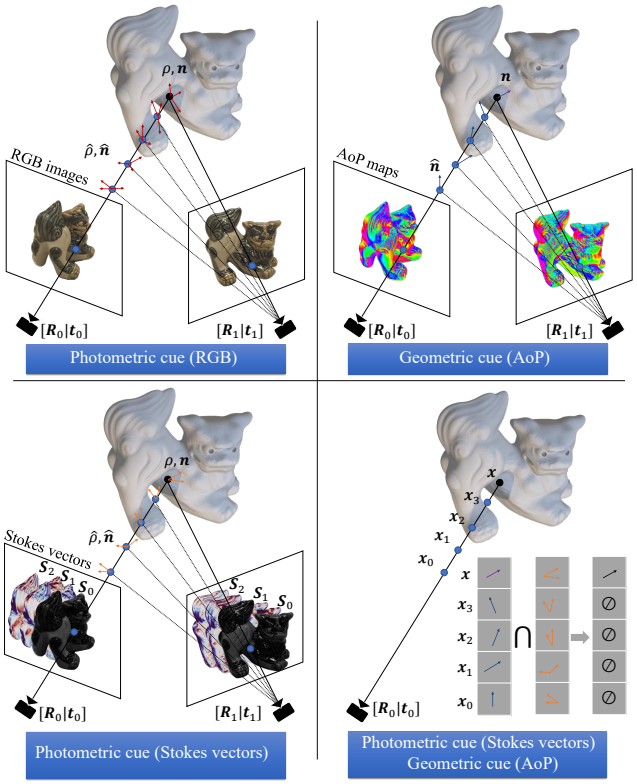

Figure 3. Ambiguity of determining 3D postions under sparse views with geometric and photometric cues.

Our method combines these two cues derived from polarized images. As visualized in the bottom-right part of Fig. 3, the correct scene point position should have its surface normal lay in the intersection of normal candidate groups derived from both photometric and geometric cues. As surface normal at different sampled scene point is uniquely determined by geometric cue, we can easily determine whether the point is on the surface with the aid of photometric cue. In this way, we reduce the solution space of sparse-shot reflective surface reconstruction.

4.2. NeRSP

Network structure As shown in Fig. 4, our NeRSP applies a similar network structure with PANDORA [9] originally derived from Ref-NeRF [30]. For a light ray emitted from camera center $\mathbf{o}$ with the direction $\mathbf{v}$, we sample a point on the ray with travel distance t_i, its location is denoted at $\mathbf{x}_i = \mathbf{o} + t_i\mathbf{v}$. Following the volume rendering used in NeRF [25], the observed Stokes vector $\mathbf{s}(\mathbf{v})$ can be integrated by the volume opacity σ_i and the Stokes vectors at the sampled points along the ray, *i.e.*

$$\mathbf{s}(\mathbf{v}) = \sum_{i=1}^{n} W_i \mathbf{s}_o(\mathbf{x}_i, \mathbf{v}) \sigma_i, \quad (11)$$

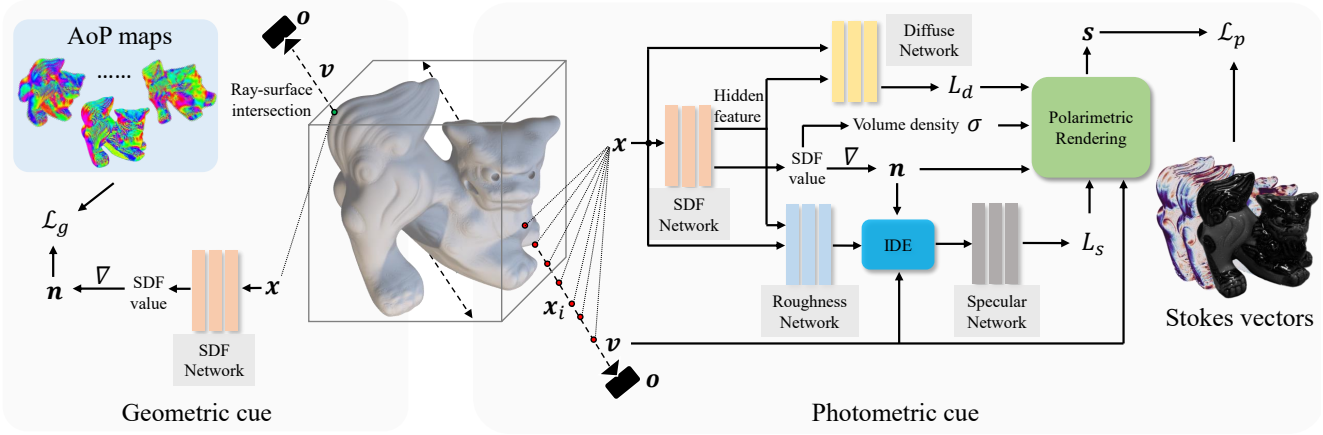

Figure 4. Pipeline of our NeRSP.

where $W_i = \prod_{j=1}^{i-1}(1-\sigma_j)$ denote the accumulated transmittance of a sampled point.

Motivated by recent neural 3D reconstruction method NeuS [33], we derives the volume opacity from a SDF network and also extract the surface normal from the gradient of the SDF. To compute $\mathbf{s}_o(\mathbf{x}_i, \mathbf{v})$ at sampled points, we follows the polarmetric image formation model in Eq. (10). Specifically, the diffuse radiance L_d is related to the diffuse albedo and Fresnel transmission coefficients, which depends on the scene positions but invariant to the view direction. Therefore, we use a diffuse radiance network to map L_d from the feature of each scene point. The specular radiance L_s is related to the specular lobe determined by the view direction, surface normal, and the surface roughness. We therefore use a RoughnessNet to predict surface roughness. Together with the camera view direction and predicted surface normal, we estimate the specular radiance L_s following the integrated positional encoding module proposed by Ref-NeRF [30]. Combining L_d and L_s, we reconstruct the observed Stokes vector following Eq. (10).

Loss function The photometric loss is defined as the L1 distance between the observed $\hat{\mathbf{s}}(\mathbf{v})$ and reconstructed Stokes vectors $\mathbf{s}(\mathbf{v})$, *i.e.*,

$$\mathcal{L}_p = \sum_{\mathbf{v}\in\mathcal{V}} \|\mathbf{s}(\mathbf{v}) - \hat{\mathbf{s}}(\mathbf{v})\|_1, \tag{12}$$

where $\mathcal{V}$ denotes all the camera rays casted within object masks at different views.

For the geometric loss. we first find the 3D scene point $\mathbf{x}$ along the camera ray $\mathbf{v}$ until touching the surface and then locate the projected 2D pixel positions at different views. The geometric loss is defined based on the Eq. (6), *i.e.*,

$$\mathcal{L}_g = \sum_{\mathbf{x}\in\mathcal{X}} \|\mathbf{T}(\mathbf{x})\mathbf{n}(\mathbf{x})\|_2^2, \tag{13}$$

where $\mathcal{X}$ denotes all the ray-surface intersections inside the object masks at different views.

Besides the photometric and geometric loss, we add mask loss supervised by the object masks and the Eikonal regularization loss. The mask loss is defined as

$$\mathcal{L}_m = \sum \mathrm{BCE}(M_k, O_k), \tag{14}$$

where $O_k = \sum_{i=1}^{n} W_{k,i}\sigma_{k,i}$ represents the predicted mask at k-th camera ray, whose GT mask value is denoted as M_k. BCE represents binary cross entropy loss.

The Eikonal loss is defined as

$$\mathcal{L}_e = \frac{1}{nf}\sum_{i,k}(\|\mathbf{n}_{i,k}\| - 1)^2, \tag{15}$$

where $\mathbf{n}_{i,k}$ is the surface normal derived from the SDF network at i-th sampled point along k-th camera ray.

Our NeRSP is supervised by the combination of the above loss terms, *i.e.*

$$\mathcal{L} = \mathcal{L}_p + \lambda_g\mathcal{L}_g + \lambda_m\mathcal{L}_m + \lambda_e\mathcal{L}_e, \tag{16}$$

where λ_e, λ_m and λ_p are the coefficients for the corresponding loss terms.

4.3. RMVP3D Dataset

To quantitatively evaluate the proposed method, we capture a Real-world Multiview Polarized image dataset with aligned ground truth meshes. Figure 5 (left) illustrates our capturing setup, which includes a polarimetric camera, FLIR BFS-U3-51S5PC-C, equipped with a 12 mm lens and a rotation rail. We use OpenCV for demosaicing the raw data and obtain 1224×1024 color images with polarizer angles at 0, 45, 90, and 135 degrees. During the data capture, we place target objects at the center of the rail, and capture 60 images per object by manually moving the camera. We collect 4 objects as targets: DOG, FROG, LION, and BALL,

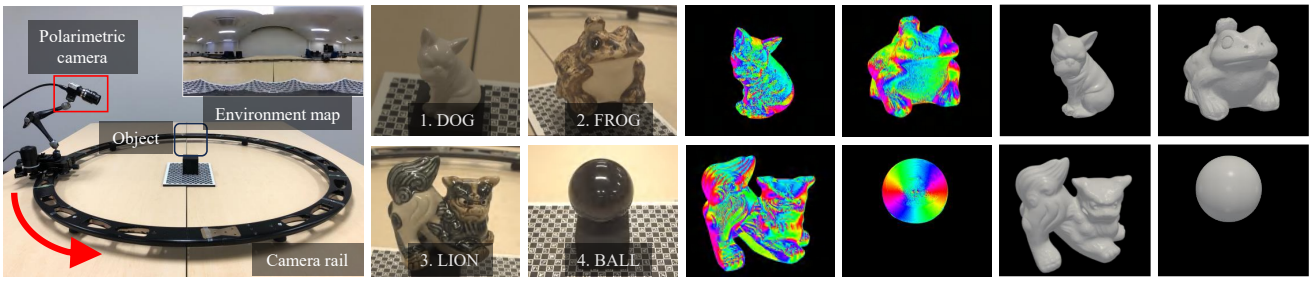

Figure 5. Capture setup and overview of our real-world multiview polarized image dataset RMVP3D.

Figure 6. Overview of synthetic dataset SMVP3D. Top and bottom rows show image observations and the corresponding AoP maps.

as shown in Fig. 5 (middle). For the quantitative evaluation, we adopt a laser scanner Creaform HandySCAN BLACK with the accuracy of 0.01 mm to obtain the ground truth mesh. To align the mesh to the captured image views, we first apply PANDORA [9] to estimated a reference shape using all available views and then align the scanned mesh to the estimated one via ICP algorithm [4]. Besides the ground-truth shapes and multiview images, we also capture the environment map using a 360-degree camera THETA Z1, benefiting quantitative evaluations on the illumination estimation for related neural inverse rendering works.

5. Experiments

We evaluate NeRSP with three experiments: 1) comparison with existing multiview 3D reconstruction methods quantitatively on synthetic dataset; 2) ablation study on the contribution of geometric and photometric loss terms 3) qualitative and quantitative evaluations on real-world datasets. We also provide the BRDF and novel view results in the supplementary material.

5.1. Datasets & Baselines

Dataset. We prepare two real-world datasets: the PANDORA dataset [9] and our proposed RMVP3D, where PANDORA dataset [9] is only used for qualitative evaluation as the ground truth meshes are not provided. We also prepare a synthetic multiview polarized image dataset SMVP3D with Mitsuba rendering engine [15], which contains 5 objects with spatially-varying and reflective reflectance, as visualized in Fig. 6. The objects are illuminated by environment maps[2] and captured by 6 views randomly distributed around the objects. Besides rendered polarized images, we also export the stokes vectors, GT surface normal maps, and AoP maps for each object.

Baselines. Our work solves multiview 3D reconstruction for reflective surfaces based on sparse polarized images. Therefore, we choose the state-of-the-art 3D reconstruction methods targeting reflective surfaces NeRO [19] and sparse views S-VolSDF [35]. The above two methods are based on RGB image inputs. For multiview stereo based on polarized images, we select PANDORA [9] and MVAS [6] as our baselines. NeRO [19] does not require silhouette masks as input. For a fair comparison, we remove the background in the RGB images with the corresponding masks before inputting to NeRO [19]. To compare different methods, we apply Chamfer distance (CD) between the estimated and the GT meshes, and the mean angular error (MAE) between the estimated and the GT surface normals at different views as our evaluation metrics.

5.2. Shape recovery on synthetic dataset

As shown in Table 1, we summarize the shape estimation error of existing methods and ours on SMVP3D. Our method achieves the smallest Chamfer distance along all of the 5 synthetic objects. Based on the visualized shape estimates shown in Fig. 7, NeRO [19] and S-VolSDF [35] cannot accurately recover surface details as highlighted in the closed-up views. One possible reason is that the disentanglement of the shape and reflective reflectance from the sparse images is too challenging for these methods based on only RGB information. MVAS [6] and PANDORA [9] address the geometric and photometric cues of the polarized images, separately. However, the reconstructed reflective surface shapes are still unsatisfactory due to the ambiguities in geometric and photometric cues under the sparse views setting. As highlighted in the closed-up views, benefiting from both geometric and photometric cues, our method reduces the

[2]https://polyhaven.com. Retrieved March, 2024.

Table 1. Comparison on shape recoveries on synthetic dataset evaluated by Chamfer distance (↓). The smallest and second smallest errors are labeled in bold and underlined. "N/A" denotes the experiment where a specific method cannot output reasonable shape estimation results.

Method	HEDGEHOG	SQUIRREL	SNAIL	DAVID	DRAGON
NeRO [19]	5.39	4.69	14.19	45.8	6.51
S-VolSDF [35]	7.33	<u>5.33</u>	16.8	<u>5.12</u>	N/A
MVAS [6]	<u>5.37</u>	5.72	<u>8.01</u>	7.01	<u>6.48</u>
PANDORA [9]	9.33	11.1	18.8	7.86	17.4
NeRSP (Ours)	**3.43**	**4.55**	**5.59**	**4.16**	**4.77**

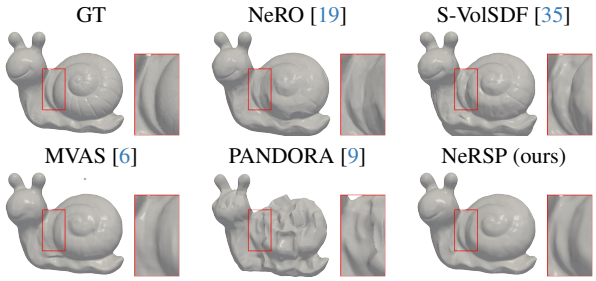

Figure 7. Qualitative evaluation on shape recoveries from 6 sparse inputs. Our recovered shapes are closer to the GT, as highlighted in the closed-up views.

solution space of shape estimation, leading to the most reasonable shape recoveries compared with the GT shapes.

Besides the evaluation of the reconstructed mesh, we also test the surface normal estimation results. As shown in Table 2, we summarize the mean angular errors of estimated surface normals at 6 views from different methods. Consistent with the evaluation results in Table 1, NeRSP achieves the smallest mean angular errors in average. We also observed that the results from NeRO [19], MVAS [6], and PANDORA [9] have larger errors on objects with fine details, such as DAVID and DRAGON objects. As an example, MVAS [6] has the second smallest Chamfer distance shown in Table 1, but the mean angular error is over 20°. One potential reason is existing methods output smooth shapes in the sparse views setting, where the surface details such as the flakes of the DRAGON are not well recovered.

5.3. Ablation study

In this section, we conduct an ablation study to test the effectiveness of geometric and photometric cues. Taking the DRAGON object as an example, we conduct our method with and without the photometric loss $\mathcal{L}_p$ and the geometric loss $\mathcal{L}_g$. As shown in Fig. 8, we plot the shape and surface normal estimations by disabling the different loss terms. Without the photometric loss, the shape ambiguity due to the sparse views occurs. As shown from the closed-up views, the shape near the leg part has a concave artifact, as there are only two visible views for this region, unable to

Table 2. Comparison on surface normal estimation on synthetic dataset evaluated by mean angular error (MAE) (↓).

Method	HEDGEHOG	SQUIRREL	SNAIL	DAVID	DRAGON
NeRO [19]	9.14	<u>10.15</u>	11.45	42.02	<u>24.22</u>
S-VolSDF [35]	11.26	13.28	7.59	<u>17.05</u>	N/A
MVAS [6]	**7.06**	10.28	<u>6.19</u>	21.86	24.29
PANDORA [9]	19.75	23.52	16.54	21.88	28.82
NeRSP (Ours)	<u>7.89</u>	**9.80**	**4.82**	**13.70**	**18.03**

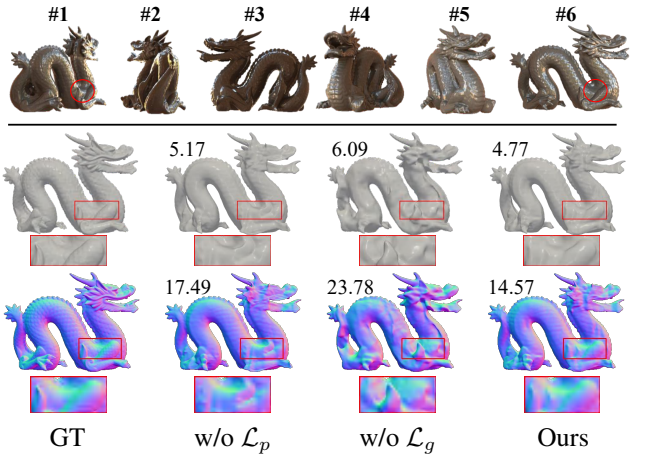

Figure 8. Ablation study on different loss terms. The top and bottom rows visualize the estimated shape and surface normal, with the Chamfer distance and the mean angular error labeled on the top of each sub-figure, respectively.

formulate a unique solution for the shape merely based on the AoP maps [6]. Without geometric loss, we also obtain distorted shape results as the sparse image observations are not sufficient to uniquely decompose the shape, reflectance, and illumination. By combining the photometric and geometric loss, our NeRSP reduces the ambiguity of shape recovery and the estimated shape is closer to the GT, as highlighted in the closed-up views.

5.4. Shape recovery on real data

Besides the synthetic experiments shown in the previous section, we also evaluate our method on real-world datasets PANDORA dataset [9] and RMVP3D to test its applicability in real-world 3D reconstruction scenarios.

Qualitative evaluation on PANDORA dataset [9]. As shown in Fig. 9, we provide qualitative evaluations on PANDORA dataset [9]. Compared to the image appearance with the estimated results from S-VolSDF [35] and NeRO [19], the shape is not fully disentangled from the reflectance, leading to bumpy surface shapes that are closely related to the reflectance texture. MVAS [6] and PANDORA [9] have over-smoothed shape estimates or concave shape artifacts, due to addressing only geometric or photometric cues un-

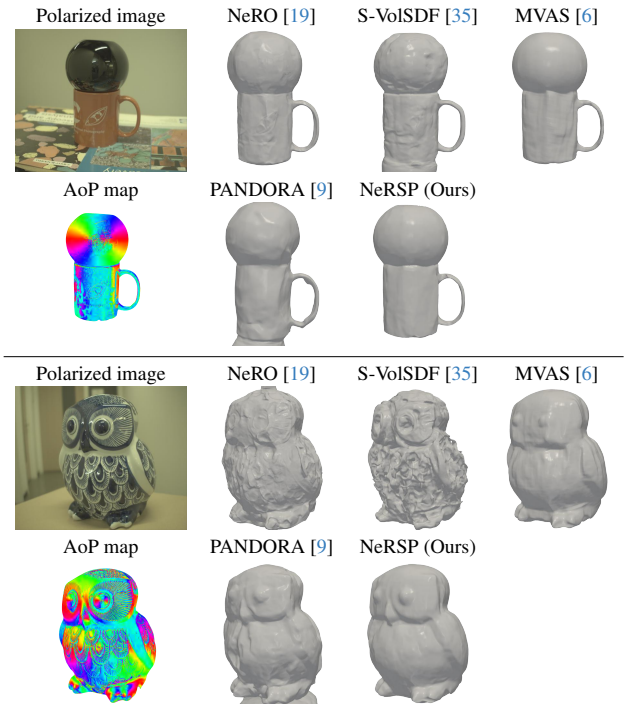

Figure 9. Qualitative evaluation on PANDORA dataset [9].

Table 3. Quantitative evaluation on RMVP3D with Chamfer distance (↓). Our method achieves the smallest error on average.

Method	DOG	LION	FROG	BALL	Average
NeRO [19]	9.11	10.74	**6.21**	3.87	7.48
S-VolSDF [35]	9.93	7.39	7.91	18.4	10.91
MVAS [6]	9.23	7.51	9.90	4.77	7.86
PANDORA [9]	14.3	15.04	11.27	3.96	11.14
NeRSP (Ours)	**8.80**	**5.18**	6.70	**3.84**	**6.13**

der the sparse capture setting. Our shape estimation results have no such shape artifacts and match the image observations closely.

Quantitative evaluation on RMVP3D. As shown in Table 3, we present a quantitative evaluation on RMVP3D based on Chamfer distance. Consistent with the synthetic experiment, our NeRSP achieves the smallest estimation error in average. The visualized shapes shown in Fig. 10 further reveal that reflective surfaces are challenging to S-VolSDF [35] for disentangling the shape from reflectance, as highlighted by the bumpy surface of the FROG object in the closed-up views. NeRO [19] and PANDORA [9] have similar estimation error with us on the simple BALL object. For complex shapes like LION, distorted shape recoveries are obtained from these methods due to the sparse views setting, while ours are closer to the GT meshes, demonstrating the effectiveness of our method on real-world reflective surface reconstruction under sparse inputs.

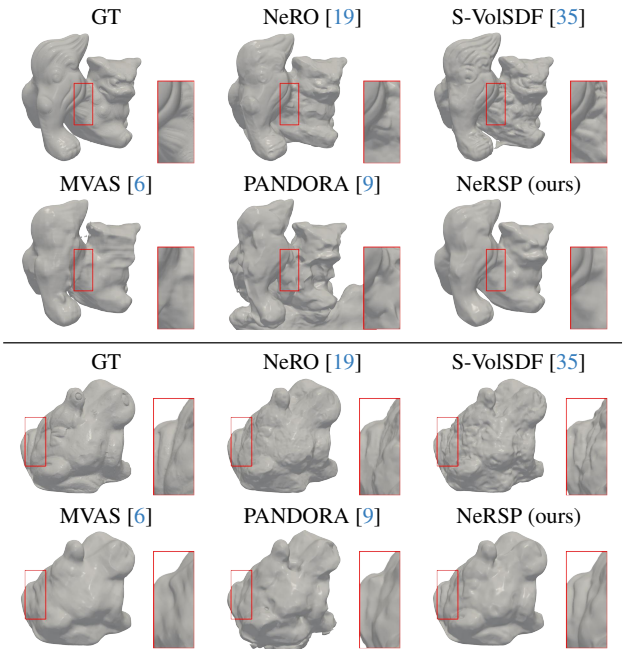

Figure 10. Comparison on shape recoveries on RMVP3D dataset.

6. Conclusion

We propose NeRSP, a neural 3D reconstruction method for reflective surfaces under sparse polarized images. Due to the challenges of shape-radiance ambiguity and complex reflectance, existing methods struggle with either reflective surfaces or sparse views and cannot address both problems with RGB images. We propose to use polarized images as input. By combining the geometric and photometric cues extracted from polarized images, we reduce the solution space of the estimated shape, allowing for the effective recovery of reflective surface with as few as 6 views, as demonstrated by publicly available and our own datasets.

Limitation The inter-reflections and polarized environment light are not considered in this work, which could influence the shape reconstruction accuracy. We noticed a most recent work NeISF [17] focusing on this topic, and we are interested in combining our sparse shot merit with this work in the future.

Acknowledgment This work was supported by the Beijing Natural Science Foundation Project No. Z200002, the National Nature Science Foundation of China (Grant No. 62136001, 62088102, 62225601, U23B2052), the Youth Innovative Research Team of BUPT No. 2023QNTD02, and the JSPS KAKENHI (Grant No. JP22K17910 and JP23H05491). We thank Youwei Lyu for insightful discussions.

References

[1] Yunhao Ba, Alex Gilbert, Franklin Wang, Jinfa Yang, Rui Chen, Yiqin Wang, Lei Yan, Boxin Shi, and Achuta Kadambi. Deep shape from polarization. In *ECCV*, pages 554–571, 2020. 2

[2] 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, 3, 4

[3] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields. In *ICCV*, pages 5855–5864, 2021. 2

[4] Paul J Besl and Neil D McKay. Method for registration of 3-D shapes. In *Sensor fusion IV: control paradigms and data structures*, pages 586–606, 1992. 6

[5] Mark Boss, Varun Jampani, Raphael Braun, Ce Liu, Jonathan Barron, and Hendrik Lensch. Neural-PIL: Neural pre-integrated lighting for reflectance decomposition. In *NeurIPS*, pages 10691–10704, 2021. 1, 2

[6] Xu Cao, Hiroaki Santo, Fumio Okura, and Yasuyuki Matsushita. Multi-View Azimuth Stereo via Tangent Space Consistency. In *CVPR*, pages 825–834, 2023. 2, 3, 4, 6, 7, 8

[7] 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 *CVPR*, pages 14124–14133, 2021. 2

[8] Zhaopeng Cui, Jinwei Gu, Boxin Shi, Ping Tan, and Jan Kautz. Polarimetric multi-view stereo. In *CVPR*, pages 1558–1567, 2017. 2

[9] Akshat Dave, Yongyi Zhao, and Ashok Veeraraghavan. Pandora: Polarization-aided neural decomposition of radiance. In *ECCV*, pages 538–556, 2022. 1, 2, 4, 6, 7, 8

[10] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised NeRF: Fewer views and faster training for free. In *CVPR*, pages 12882–12891, 2022. 2

[11] Yuqi Ding, Yu Ji, Mingyuan Zhou, Sing Bing Kang, and Jinwei Ye. Polarimetric helmholtz stereopsis. In *ICCV*, pages 5037–5046, 2021. 2

[12] Yoshiki Fukao, Ryo Kawahara, Shohei Nobuhara, and Ko Nishino. Polarimetric normal stereo. In *CVPR*, pages 682–690, 2021. 2

[13] Wenhang Ge, Tao Hu, Haoyu Zhao, Shu Liu, and Ying-Cong Chen. Ref-NeuS: Ambiguity-Reduced Neural Implicit Surface Learning for Multi-View Reconstruction with Reflection. *arXiv preprint arXiv:2303.10840*, 2023. 1

[14] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 1

[15] Wenzel Jakob. Mitsuba renderer, 2010. 6

[16] Chenyang Lei, Chenyang Qi, Jiaxin Xie, Na Fan, Vladlen Koltun, and Qifeng Chen. Shape from polarization for complex scenes in the wild. In *CVPR*, pages 12632–12641, 2022. 2

[17] Chenhao Li, Taishi Ono, Takeshi Uemori, Hajime Mihara, Alexander Gatto, Hajime Nagahara, and Yuseke Moriuchi. NeISF: Neural Incident Stokes Field for Geometry and Material Estimation. *arXiv preprint arXiv:2311.13187*, 2023. 8

[18] Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. Neuralangelo: High-Fidelity Neural Surface Reconstruction. In *CVPR*, pages 8456–8465, 2023. 2

[19] 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. 1, 2, 4, 6, 7, 8

[20] Xiaoxiao Long, Cheng Lin, Peng Wang, Taku Komura, and Wenping Wang. SparseNeuS: Fast generalizable neural surface reconstruction from sparse views. In *ECCV*, pages 210–227, 2022. 2

[21] Youwei Lyu, Lingran Zhao, Si Li, and Boxin Shi. Shape from polarization with distant lighting estimation. *IEEE TPAMI*, 2023. 2

[22] 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 *ECCV*, pages 405–421, 2020. 1, 2

[23] Miyazaki, Tan, Hara, and Ikeuchi. Polarization-based inverse rendering from a single view. In *ICCV*, pages 982–987, 2003. 2

[24] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3D representations without 3D supervision. In *CVPR*, pages 3504–3515, 2020. 2

[25] 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 *CVPR*, pages 5480–5490, 2022. 2, 4

[26] Michael Oechsle, Songyou Peng, and Andreas Geiger. UNISURF: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *ICCV*, pages 5589–5599, 2021. 2

[27] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *CVPR*, pages 165–174, 2019. 1

[28] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *NeurIPS*, 2020. 1

[29] William AP Smith, Ravi Ramamoorthi, and Silvia Tozza. Height-from-polarisation with unknown lighting or albedo. *IEEE TPAMI*, 41(12):2875–2888, 2018. 2

[30] 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. 4, 5

[31] 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. 4

[32] 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

[33] 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, 2, 5

[34] Yiqun Wang, Ivan Skorokhodov, and Peter Wonka. HF-NeuS: Improved surface reconstruction using high-frequency details. In *NeurIPS*, pages 1966–1978, 2022. 2

[35] Haoyu Wu, Alexandros Graikos, and Dimitris Samaras. S-VolSDF: Sparse Multi-View Stereo Regularization of Neural Implicit Surfaces. *arXiv preprint arXiv:2303.17712*, 2023. 1, 2, 4, 6, 7, 8

[36] Jiawei Yang, Marco Pavone, and Yue Wang. FreeNeRF: Improving Few-shot Neural Rendering with Free Frequency Regularization. In *CVPR*, pages 8254–8263, 2023. 2

[37] 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 *NeurIPS*, pages 2492–2502, 2020. 2

[38] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *NeurIPS*, pages 4805–4815, 2021. 1, 2

[39] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, pages 4578–4587, 2021. 2

[40] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. NeRF++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 2

[41] Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. PhySG: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *CVPR*, pages 5453–5462, 2021. 1, 2

[42] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. NeRFactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM TOG*, 40(6):1–18, 2021. 2

[43] Jinyu Zhao, Yusuke Monno, and Masatoshi Okutomi. Polarimetric multi-view inverse rendering. *IEEE TPAMI*, 2022. 2

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

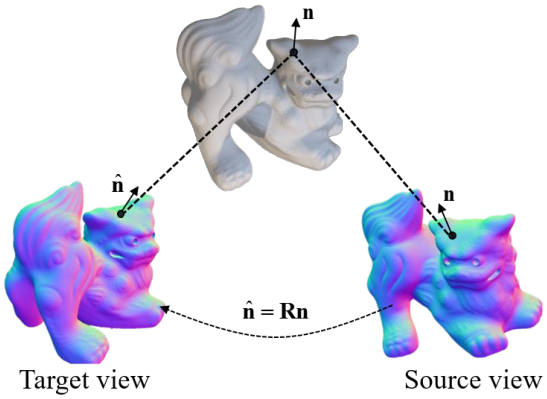

Figure S1. A scene point observed by the target view and the source view.

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

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

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

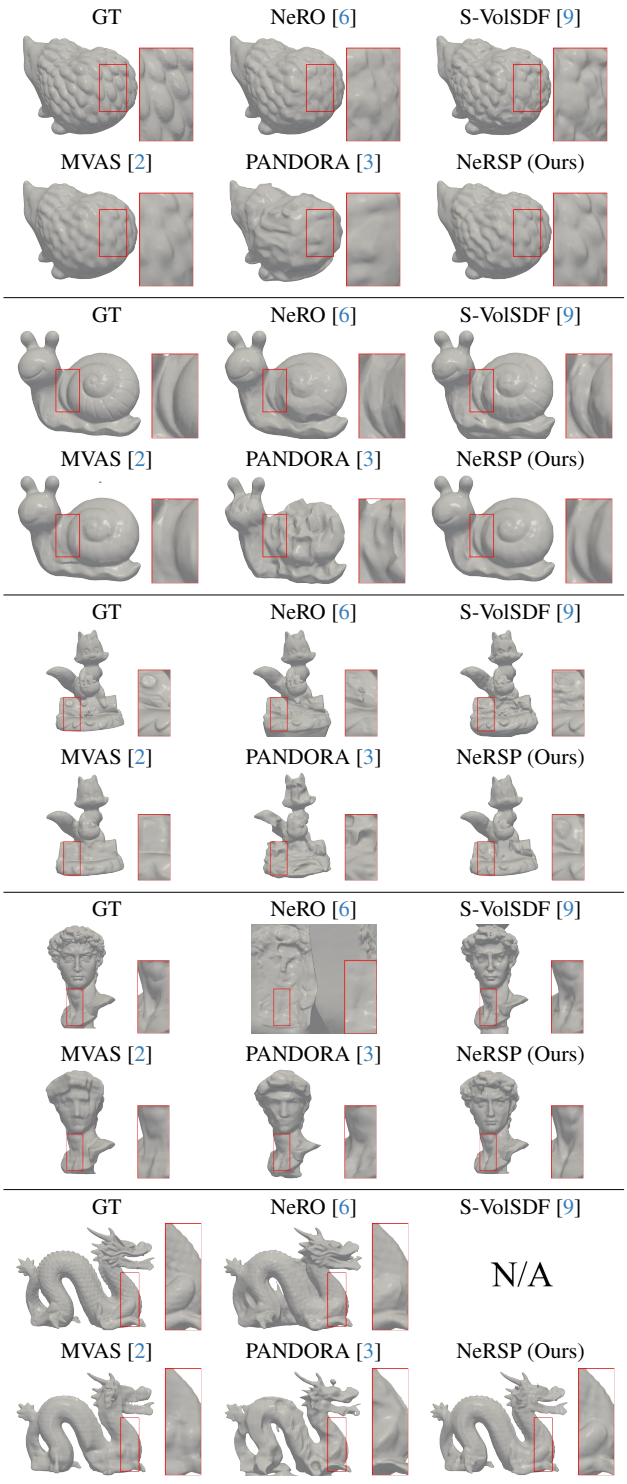

Figure S2. Qualitative evaluation of shape reconstruction on SMVP3D.

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,

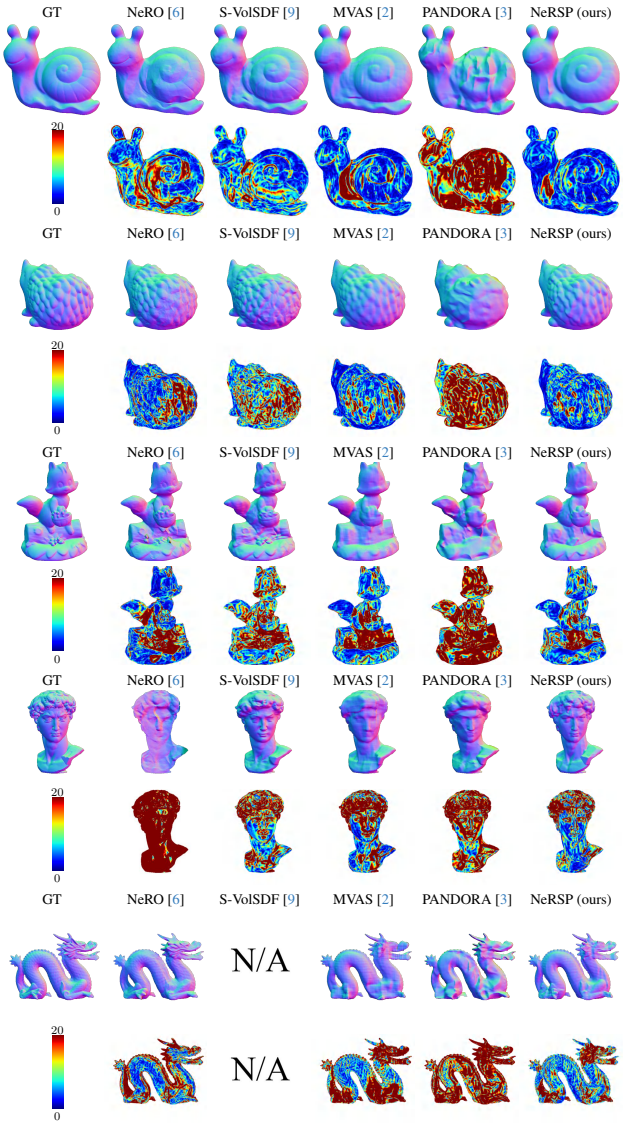

Figure S3. Qualitative evaluation of surface normal estimation on our SMVP3D. Even and odd rows show the surface normal estimates and the corresponding angular error distributions.

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.

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

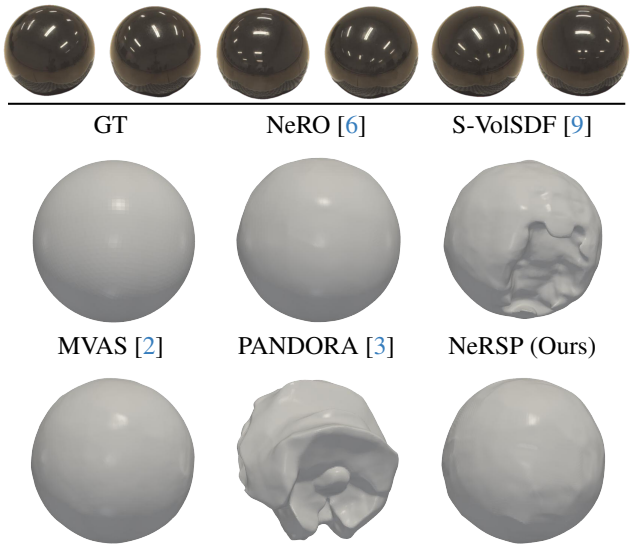

Figure S5. Qualitative evaluation of shape reconstruction BALL.

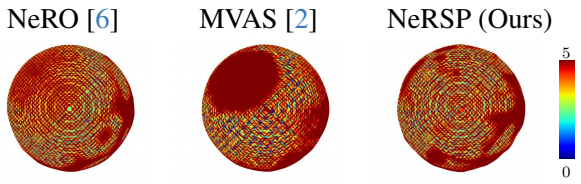

Figure S6. The Chamfer Distance maps clipped between 0 and $5\,mm$ for the estimated shapes of BALL from NeRO [6], MVAS [2], and NeRSP.

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**

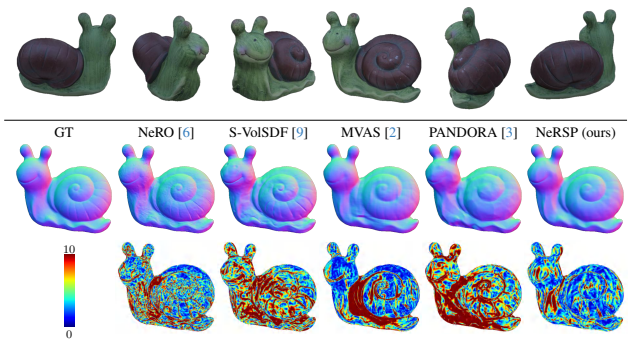

Figure S7. Qualitative evaluation of surface normal estimation on SNAIL with less reflective reflectance. Top row shows the 6 input views. Second and third rows show the surface normal estimates and the corresponding angular error distributions.

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

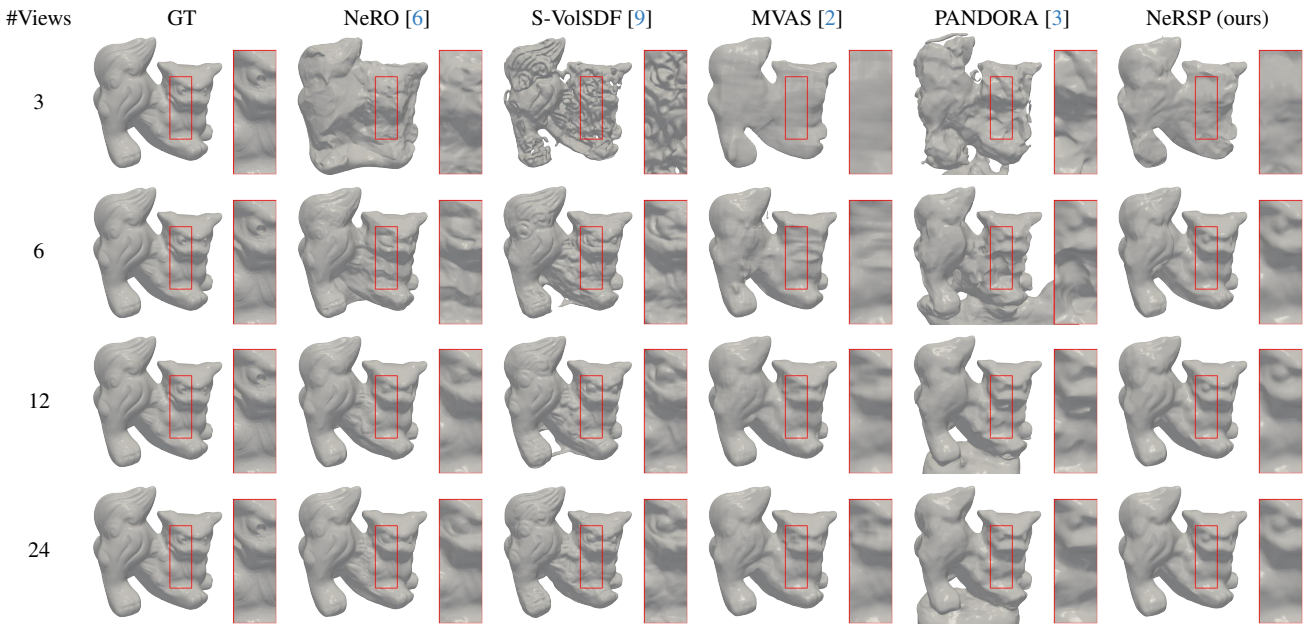

Figure S8. Qualitative results of shape reconstruction on LION with different input #views.

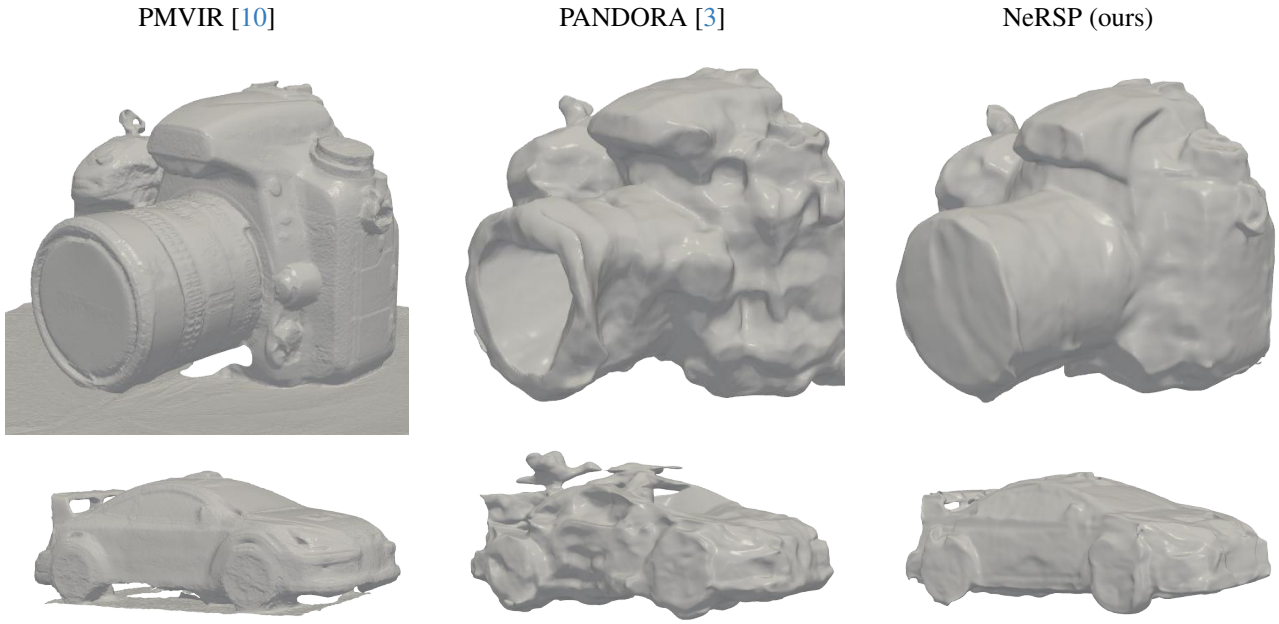

Figure S9. Shape estimation results on the Polarimetric MVIR dataset [10]. PANDORA [3] and our NeRSP use polarized images with 6 sparse views as input. As a reference, PMVIR [10] uses 31 and 56 input views on the camera and the car cases, respectively.