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

Mingzhi Pei<sup>1</sup> Xu Cao<sup>2</sup> Xiangyi Wang<sup>1</sup> Heng Guo<sup>1\*</sup> Zhanyu Ma<sup>1</sup> <sup>1</sup>Beijing University of Posts and Telecommunications <sup>2</sup>Independent Researcher <sup>{pmz, e1777216wxy, guoheng, mazhanyu}@bupt.edu.cn xucao.42@gmail.com</sup>

### Abstract

Reflective and textureless surfaces remain a challenge in multi-view 3D reconstruction. Both camera pose calibration and shape reconstruction often fail due to insufficient or unreliable cross-view visual features. To address these issues, we present PMNI (Pose-free Multi-view Normal Integration), a neural surface reconstruction method that incorporates rich geometric information by leveraging surface normal maps instead of RGB images. By enforcing geometric constraints from surface normals and multi-view shape consistency within a neural signed distance function (SDF) optimization framework, PMNI simultaneously recovers accurate camera poses and high-fidelity surface geometry. Experimental results on synthetic and real-world datasets show that our method achieves state-of-the-art performance in the reconstruction of reflective surfaces, even without reliable initial camera poses.

# 1. Introduction

Detailed 3D reconstruction from multi-view image observations is a fundamental task in computer vision, empowering various applications like virtual reality and e-Heritage. A typical pipeline first calibrates the camera poses for each image and then uses the posed images to recover the shape. Many methods have achieved promising results in scenes with diffuse and specular surfaces [2, 9, 10, 13, 15, 18, 24]. However, surface reconstruction without pose calibration, which is desired for practical causal capture scenarios, still remains challenging for reflective and textureless surfaces, as shown in Fig. 1.

Existing methods for reconstructing reflective and textureless surfaces, such as NeRO [18], require precise camera pose calibration. To achieve this, a calibration board is often placed under the object, limiting the method's applicability in more casual setups. However, jointly recovering



Figure 1. (**Top row**) Given multi-view surface normals of a reflective and textureless surface, our method jointly recovers a highfidelity surface (**middle row**) and accurate camera poses (**bottom row**). The reconstructed shape is comparable to the results of [3], which uses calibrated poses.

camera poses and surface from images presents a chickenand-egg problem. As shown in Fig. 2, without knowing the relative pose between cameras  $\mathbf{c}_1$  and  $\mathbf{c}_2$ , the epipolar plane  $\mathbf{c}_1 - \mathbf{c}_2 - \mathbf{x}$  remains ambiguous. To mitigate this ambiguity, existing methods either attempt to establish feature correspondences  $[\mathbf{p}_1, \mathbf{p}_2]$  [1, 17] between views or to constrain the shape at specific locations using monocular depth estimation [8, 25]. However, establishing reliable feature correspondences for reflective surfaces is particularly challenging due to view-dependent reflectance. Moreover, the lack of texture further complicates shape estimation using learning-based monocular depth estimators [20, 26]. Consequently, there remains a need for a robust 3D reconstruc-

<sup>\*</sup> Corresponding author.

Released at: https://github.com/pmz-enterprise/PMNI



Figure 2. Illustration of shape and pose estimation for reflective and textureless objects based on RGB and surface normals.

tion method that can accurately handle reflective and textureless surfaces while being tolerant to noisy camera poses.

In this paper, we propose Pose-free Multi-view Normal Integration (PMNI), a method that leverages multi-view surface normal maps as input to jointly optimize both surface shape and camera poses. Our key insight is that monocular normal estimation is independent of camera poses, and normal maps encode rich shape information that aids in camera pose estimation. As shown in Fig. 2, a normal map can be estimated by photometric stereo from singleview image observations [11], and is robust to textureless and reflective surfaces. By applying the normal integration method [4], relative depth maps with fine-grained details can be recovered from single-view surface normal maps, providing reliable geometric cues that facilitate camera pose estimation. Moreover, unlike RGB images, where photometric consistency is often disrupted for reflective surfaces, surface normals remain geometrically consistent at corresponding points, making them invariant to changes in surface reflectance.

Building on these insights, we propose a pose-free reflective surface reconstruction method based on multi-view surface normal maps. Specifically, we utilize a signed distance function (SDF) represented by a coordinate-based MLP network, which can simultaneously model both surface shape and surface normals through its analytical gradient. We use per-view depth map, integrated from the surface normal map [4], as an anchor to regularize the SDF. At each iteration, with the estimated shape and camera poses, we find correspondences by projecting sampled rays onto the image plane. This allows us to further constrain both the shape and poses by enforcing the geometric consistency of the input surface normals at their projected positions.

As shown in Fig. 1, PMNI enables the joint recovery of high-fidelity 3D shapes and camera poses, yielding results comparable to methods with calibrated camera poses [3], and outperforming existing pose-free 3D reconstruction approaches [25]. In this way, PMNI makes it possible for re-

	Camera pose calibration	<u>_</u>	Pose calibration free	Õ
Diffuse & textured	<ul> <li>NueS (RGB, SDF)</li> <li>VolSDF (RGB, SDF)</li> <li>IDR (RGB, SDF)</li> <li>Gaussian Surfels (R</li> </ul>	) 1GB, 3DGS)	<ul> <li>COLMAP (RGB, F</li> <li>Nope-NeRF (RGB, SPARF (RGB, NeR</li> <li>DUS43R (RGB, Poil</li> <li>CF-3DGS (RGB, 31</li> <li>COGS (RGB, 3DG</li> </ul>	Point cloud) NeRF) F) nt cloud) DGS) S)
Reflective & textureless	<ul> <li>NeRO (RGB, SDF)</li> <li>Ref-NeuS (RGB, SE</li> <li>PISR (Polarization in</li> <li>PANDORA (Polarization</li> <li>NeRSP (Polarization</li> <li>NeISF (Polarization</li> <li>SuperNormal (Surfa</li> <li>RnbNeuS (MVML in</li> </ul>	DF) mage, SDF) ation image, SDF) n image, SDF) image, SDF) ace normal, SDF) mages, SDF)	Ours	

Figure 3. Summary of existing neural surface reconstruction methods categorized by their surface reflectance types and camera calibration settings. The input and surface representation for each method are labeled in brackets.

flective and textureless surface reconstruction in a causal capture setting.

**Contributions.** This paper proposes PMNI, the first method to achieve high-quality reflective surface reconstruction without camera pose calibration. Unlike RGB images, surface normal maps derived from photometric stereo are invariant to reflective and textureless surfaces. We demonstrate normal maps provide an effective regularization for surface reconstruction through integrated depth, effectively reducing ambiguities in both shape and pose recovery. Experiments on both public and our own captured real-world datasets validate the effectiveness of our method.

## 2. Related works

This paper focuses on pose-free reflective surface reconstruction from multi-view surface normal maps. In the following, Sec. 2.1 summarizes related works on reflective surface reconstruction that take image observations or surface normal maps as input. Section 2.2 then surveys pose-free neural radiance field (NeRF) methods.

### 2.1. Neural reflective surface reconstruction

Neural 3D reconstruction has advanced significantly since NeRF [19]. Given multi-view images, camera poses are estimated via structure-from-motion through feature matching. Shape represented by SDF [24] or Gaussian surfels [6] is then optimized with differentiable volume rendering.

Reflective surfaces pose additional challenges due to view-dependent reflections, as shown in Fig. 3. Methods like NeRO [18] and Ref-NeuS [9] effectively address these issues by using RGB inputs and incorporating Integrated Positional Encoding (IDE) and split-sum approximations to model reflective appearance under environmental lighting. Polarization-based neural reconstruction, such as PAN-DORA [7], NeRSP [10], and PISR [5], uses the polarization characteristics of diffuse and specular reflectance to address shape-reflectance ambiguity. By decomposing radiance into diffuse and specular components via Stokes vector, these methods improve reflective surface reconstruction. NeRSP [10]further integrates geometric cues from the angle of polarization with photometric cues from Stokes vectors, enabling shape estimation even with sparse views.

Photometric stereo excels in reconstructing single-view shapes with complex reflectance by taking images under varying lighting as input. Multi-view photometric stereo (MVPS) extends this approach by combining multi-view, multi-light observations. Methods like SuperNormal [3] and RNb-Neus [2] first extract per-view normal maps using techniques like SDM-UniPS [11], then refine SDF to align with these normals. Compared to RGB- or polarizationbased methods, MVPS is particularly effective for reflective surfaces due to the detailed geometric information encoded in surface normal maps [3].

As shown in Table 3, feature extraction for reflective and textureless surfaces is highly challenging, making camera pose calibration with COLMAP [21] unreliable. Consequently, existing reflective surface reconstruction methods rely on a checkerboard during capture, which limits their applicability in casual capture settings. In contrast, PMNI leverages multi-view surface normal maps as input to achieve detailed reflective surface reconstruction without requiring pose calibration.

### 2.2. Pose-free surface reconstruction

Given multi-view images, COLMAP [21] uses Structure from Motion (SfM) to reconstruct camera poses and sparse 3D points from feature correspondences. For reflective surfaces, where feature matching is challenging, adding a checkerboard can improve reliability.

To address pose errors in COLMAP [21], pose-free methods have been developed to jointly recover surface shapes and camera poses. As summarized in Table 3, BARF [17] optimizes poses and NeRF using coarse-to-fine positional encoding but requires pose initializations close to the ground truth. CF-3DGS [8] mitigates this by enforcing temporal continuity and using explicit point cloud representations, though it is limited to dense video sequences.

Pose-free shape reconstruction from sparse views often incorporates learning-based priors. SPARF [22] and DUSt3R [25] rely on pre-trained networks to establish dense 2D correspondences or 2D-to-3D mappings. COGS [12] and Nope-NeRF [1] use monocular depth estimators (*e.g.*, Marigold [14] and DPT [20]) to assist in shape estimation without camera poses. However, these approaches struggle with reflective surfaces.

In summary, pose-free methods often depend on spe-

cific initializations, pose continuity, or learning-based priors, which are less effective for reflective surfaces. Additionally, NeRF-based or 3D Gaussian Splatting-based methods often yield noisy shapes. In contrast, PMNI uses SDF representation and surface normal input to achieve detailed shape reconstruction without precise pose initialization.

# 3. Proposed method

We aim to jointly recover fine-grained shapes and camera extrinsic parameters from (1) multi-view camera-space normal images, (2) the corresponding foreground masks of the target object, and (3) camera intrinsic parameters. To this end, we first perform monocular normal integration on perview normal maps to obtain per-view relative depth maps; then the normal and depth maps are used together to guide camera pose and shape optimization.

## 3.1. Preliminaries

**SDF-based neural surface reconstruction.** Signed Distance Function (SDF) is a common implicit representation of the 3D shape. The surface of the object  $\mathcal{M}$  can be viewed as the zero-level set of the SDF:

$$\mathcal{M} = \{ \mathbf{x} \mid f(\mathbf{x}) = 0 \}. \tag{1}$$

Based on NeuS [24], implicit representation of SDF is connected with volume rendering. Specifically, given Kordered 3D points  $\{\mathbf{x}_i\}_{i=0}^{K}$  on a ray and their SDF values  $\{f(\mathbf{x}_i)\}_{i=0}^{K}$ , the volume opacity of a point in space is calculated as follows:

$$\alpha_i = \max\left(\frac{\phi_s(f(\mathbf{x}_i)) - \phi_s(f(\mathbf{x}_{i+1}))}{\phi_s(f(\mathbf{x}_i))}, 0\right), \quad (2)$$

where  $\phi_s(x) = 1/(1 + \exp(-sx))$  is the sigmoid function with a learnable sharpness s. The accumulated transmittance  $T_i$  at a point along a ray can be expressed as:

$$T_i = \prod_{j=0}^{i-1} (1 - \alpha_j).$$
 (3)

Following volume rendering, we render the depth, surface normal, and opacity of a pixel **p** by

$$\hat{z}\left(\mathbf{p}\right) = \sum_{i=0}^{K} T_i \alpha_i d_i,\tag{4}$$

$$\hat{\mathbf{n}}^{w}(\mathbf{p}) = \sum_{i=0}^{K} T_{i} \alpha_{i} \nabla f(\mathbf{x}_{i}), \qquad (5)$$

$$\hat{o}(\mathbf{p}) = \sum_{i=0}^{K} T_i \alpha_i, \tag{6}$$

where  $\nabla f(\mathbf{x})$  denotes the gradient of SDF,  $\hat{\mathbf{n}}^w$  denotes world-space surface normal. By supervising the above volume-rendering information, the SDF network can be constrained, thus accomplishing 3D reconstruction.

**Single-view normal integration** aims at reconstructing the relative height map from a given normal map. We define surface normal at **p** as  $\mathbf{n}^c = [n_x, n_y, n_z]^{\top}$  in the camera space, the gradient field  $[p, q]^{\top}$  under orthographic projection can be extracted as

$$p = -\frac{n_x}{n_z}, \quad q = -\frac{n_y}{n_z}.$$
 (7)

The normal integration problem can be formulated as minimizing the following functional:

$$\mathcal{J}(z) = \iint_{\Omega_n} \left( (\partial_u z - p)^2 + (\partial_v z - q)^2 \right) du \, dv, \quad (8)$$

where  $\partial_u$  and  $\partial_v$  denote partial derivatives of the depth function  $z : \Omega_n \to \mathbb{R}$  along u and v axes on the image plane. Based on this optimization, single-view depth can be obtained up to scale under perspective projection [4].

### 3.2. Problem definition

Given camera intrinsic K and multi-view surface normal maps in the camera space, the problem we aim to solve can be formulated as:

$$\min \|\mathbf{R}_i \nabla f(\mathbf{x}) - \mathbf{n}_i^c \left( \mathbf{K} (\mathbf{R}_i \mathbf{x} + \mathbf{t}_i) \right) \|_2^2, \qquad (9)$$

where  $[\mathbf{R}_i, \mathbf{t}_i]$  denote rotation and translation at *i*-th view,  $\mathbf{p} = \mathbf{K}(\mathbf{R}_i\mathbf{x} + \mathbf{t}_i)$  represents the projected pixel position of  $\mathbf{x}$  at the view, and  $\mathbf{n}^c(\mathbf{p})$  denotes the observed surface normal in camera space.  $\mathbf{R}_i \nabla f(\mathbf{x})$  rotates the world-space normal  $\mathbf{n}^w = \nabla f(\mathbf{x})$  to the camera space. By minimizing the difference in camera-space surface normal, this paper aims to solve the 3D surface shape represented by SDF, and the multi-view camera poses jointly.

#### 3.3. Joint optimization of pose and surface

PMNI adopts a hash-encoded SDF network and uses volume rendering to get per-view surface normal and depth in the world space. We set the SDF network parameters and camera poses as learnable variables, which are optimized via the following loss function:

$$\mathcal{L} = \lambda_0 \mathcal{L}_{normal} + \lambda_1 \mathcal{L}_{ni} + \lambda_2 \mathcal{L}_c + \lambda_3 \mathcal{L}_{eikonal} + \lambda_4 \mathcal{L}_{mask},$$
(10)

where  $\lambda_i$  is the coefficient to balance different loss terms. In the following, we introduce details of these loss terms. World-to-camera surface normal loss  $\mathcal{L}_{normal}$ . Given world-space surface normal  $\mathbf{n}^w(\mathbf{p})$  projected at pixel location  $\mathbf{p}$  rendered from SDF network, and the camera-space surface normal  $\mathbf{n}^c(\mathbf{p})$  recorded in the input surface normal map,  $\mathcal{L}_{normal}$  is defined as

$$\mathcal{L}_{normal} = \sum_{i=1}^{N} \sum_{\mathbf{p}} |\mathbf{R}_i \mathbf{n}_i^w(\mathbf{p}) - \mathbf{n}_i^c(\mathbf{p})|_2^2, \quad (11)$$

where N denotes the number of input views.

**Normal integration loss**  $\mathcal{L}_{ni}$ . Given surface normal maps, we use normal integration method BiNI [4] to get integrated depth map  $\mathbf{z}^{ni}$ . This depth map has an inherent scale ambiguity to the corresponding GT depth, *i.e.*,  $\mathbf{z} = \alpha \mathbf{z}^{ni}$ . Given the depth map  $\mathbf{z}^r$  rendered from SDF, we can calculate this scale  $\alpha$  via least squares. Specifically,

$$\alpha = \frac{\mathbf{z}^{ni} \cdot \mathbf{z}^r}{\mathbf{z}^r \cdot \mathbf{z}^r}.$$
 (12)

We calculate the scale for each view. Using these integrated depth maps, we regularize the SDF network by L1 loss, *i.e.*,

$$\mathcal{L}_{ni} = \sum_{i=1}^{N} \left| \mathbf{z}_{i}^{r} - \alpha_{i} \mathbf{z}_{i}^{ni} \right|.$$
(13)

**Multi-view normal consistency loss**  $\mathcal{L}_c$ . Motivated by existing pose-free 3D reconstruction methods that apply correspondences between views to regularize the camera poses, we try to find dynamic 2D correspondences by projecting sampled scene points from SDF to the 2D image plane defined by camera poses at each iteration. Given these 2D correspondences, we measure the consistency of the surface normal maps in the camera space.

Specifically, we first cast a ray passing through p at a reference view and trace until touching the surface points x.

After that, we project  $\mathbf{x}$  to all the other camera views via projection  $\Pi = \{\pi_i = [R_i, t_i] \mid i = 0, ..., N - 1\}$ , and get the corresponding surface normals  $\mathbf{n}^c(\pi_i(\mathbf{x}))$  in the camera space. Theoretically, these surface normals can be rotated to the same world surface normal via the corresponding camera poses. Based on this constraint, we define the loss at  $\mathbf{p}$ as

$$\mathcal{L}_{c} = \sum_{i}^{N-1} \left\| \bar{\mathbf{R}} \bar{\mathbf{n}}^{c}(\mathbf{p}) - \mathbf{R}_{i} \mathbf{n}_{i}^{c}(\pi_{i}(\mathbf{x})) \right\|_{2}^{2}, \qquad (14)$$

where  $\bar{\mathbf{R}}$  and  $\bar{\mathbf{n}}$  denote the rotation and camera view normal in the reference frame. As point  $\mathbf{x}$  is not visible to all views, we introduced a visible mask function  $\gamma_i(x)$  based on ray tracing, *i.e.*,

$$\gamma_i(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \text{ is visible to } i\text{-th camera,} \\ 0 & \text{otherwise.} \end{cases}$$
(15)

Based on this visibility check, we rewrite the multi-view normal consistency loss as

$$\mathcal{L}_{c} = \sum_{i}^{N-1} \gamma_{i}(\mathbf{x}) \left\| \bar{\mathbf{R}} \bar{\mathbf{n}}^{c}(\mathbf{p}) - \mathbf{R}_{i} \mathbf{n}_{i}^{c}(\pi_{i}(\mathbf{x})) \right\|_{2}^{2}.$$
 (16)

We compute and accumulate this loss under different pixels and reference views.

**Mask loss**  $\mathcal{L}_{mask}$  is built upon labeled silhouette of target shape, *i.e.*,

$$\mathcal{L}_{mask} = \sum_{i}^{N} \sum_{\mathbf{p}} \text{BCE}\left(\hat{o}_{i}(\mathbf{p}), o_{i}(\mathbf{p})\right), \qquad (17)$$

where  $o_i(\mathbf{p})$  and  $\hat{o}_i(\mathbf{p})$  correspond to the input and rendered mask value at  $\mathbf{p}$ , and  $BCE(\cdot)$  denotes binary cross entropy function.

**Eikonal loss**  $\mathcal{L}_{eikonal}$ . To enforce the SDF gradient norm to be close to 1 almost everywhere so that the neural SDF is approximately valid, we introduce Eikonal loss as follows,

$$\mathcal{L}_{eikonal} = \sum_{\mathbf{x}} \left( \|\nabla f(\mathbf{x})\|_2 - 1 \right)^2.$$
(18)

**Initialization of camera poses.** We initialize the camera poses as a circular distribution with a radius r. To determine r, we assume the target object is within a bounding box  $[-1, 1]^3$ , and the object is always fully covered by each view with the resolution of  $H \times W$ . Given focal length in pixel units, radius r can be determined by 2f/H. More details can be found in our supplementary material.

### 4. Experiment

In this section, we first evaluate the shape estimation of our method using multi-view normal integration with calibrated camera poses as reference (Sec. 4.1). Then, we compare pose-free 3D reconstruction methods with ours on both pose and shape estimation (Sec. 4.2). More experiments, such as shape reconstruction from sparse and uncalibrated camera poses, are in the supplementary material.

### 4.1. Comparison on multi-view normal integration

**Dataset and baselines.** DiLiGenT-MV [16] includes 5 objects captured from 20 views, providing ground-truth 3D meshes, per-view surface normals, and calibrated camera poses. Using DiLiGenT-MV [16], we compare our method with the state-of-the-art multi-view normal integration method SuperNormal [3], which uses calibrated camera poses as input.

Table 1. Quantitative evaluation of shape and camera pose recovery on DiLiGenT-MV [16]. SuperNormal [3] with noisy camera poses is indicated with \* marker. The best and second-best results are labeled in **bold** and underlined.

Method	Metric	BEAR	BUDDHA	Cow	Рот2	READING	Average
SuperNormal [3]	$CD\downarrow$	0.158	<b>0.111</b>	0.099	0.154	0.187	0.142
SuperNormal [3] *		0.614	0.862	0.985	0.771	0.645	0.775
PMNI		0.189	<u>0.122</u>	<u>0.191</u>	0.115	0.148	0.153
SuperNormal [3]	F1-score ↑	0.982	<b>0.998</b>	0.999	0.998	<u>0.988</u>	0.993
SuperNormal [3] *		0.500	0.356	0.310	0.421	0.465	0.410
PMNI		0.970	<u>0.996</u>	0.989	<b>0.999</b>	<b>0.995</b>	<u>0.990</u>
PMNI	$\begin{array}{c} \text{RPEr}(^{\circ}) \downarrow \\ \text{RPEt} \downarrow \end{array}$	0.115 0.030	0.231 0.059	0.184 0.044	0.141 0.037	0.209 0.087	0.176 0.051



Figure 4. Qualitative comparison between SuperNormal [3] (abbreviated by SN) and ours on DiLiGenT-MV [16]. The camera poses for "SN noise" are slightly perturbed to simulate calibration noise. Our method accurately recovers camera poses, and the reconstruction is robust to pose calibration noise.

**Evaluation metric.** We evaluate shape accuracy using the L2 Chamfer distance (CD) and F-score with a threshold of  $\tau_F = 0.5mm$ , following SuperNormal [3]. For pose estimation, we align the estimated poses to the ground truth using [23]. Following Nope-NeRF [1], Relative Pose Error (RPE) is adopted, consisting of relative rotation error (RPEr) and relative translation error (RPEt) to assess errors between image pairs.

Shape and pose recovery results. As shown in Table 1, we perform a quantitative comparison between SuperNormal [3] and PMNI on DiLiGenT-MV [16], using the GT surface normal as input. We observe that PMNI achieves comparable 3D shape reconstruction to SuperNormal [3], with better results on the READING object. As shown in Fig. 4, the CD error distributions show that both SuperNormal [3] and our method recover shapes close to the GT, highlighting the effectiveness of our pose-free multi-view normal integration approach.

We also evaluate SuperNormal [3]'s robustness to camera pose calibration. Specifically, Gaussian noise with variations of 0.01 for translation and 0.287° for rotation is added to the camera poses. Despite these small perturbations, the recovered shapes exhibit high-frequency artifacts, as shown in Fig. 4. This occurs because noisy camera poses lead to inconsistencies in multi-view surface normal projections. In contrast, by jointly optimizing camera poses and surface shapes, our PMNI method is robust to calibration noise and produces significantly smaller shape estimation errors than SuperNormal [3].

Evaluation under different input surface normals. Multi-view normal integration is flexible regarding the source of input surface normal maps. From a practical standpoint, it is important to assess the robustness of PMNI to errors introduced by different normal estimators. Specifically, we use the photometric stereo method SDM-UniPS [11], which relies on images under varying lighting, and the single-image normal estimator StableNormal [27] to generate input surface normals for both our method and SuperNormal [3]. As shown in Fig. 5, surface normals from StableNormal [27] are less accurate than those from SDM-UniPS [11]. However, PMNI still achieves 3D shape estimations comparable to SuperNormal [3]. Despite differences in input, the recovered poses based on StableNormal [27] or SDM-UniPS [11] remain close to the GT. Table 2 further summarizes shape and pose estimation errors on DiLiGenT-MV [16], showing that PMNI and SuperNormal [3] are comparable, which highlights the robustness of our method to varying levels of input surface normal errors.

### 4.2. Comparison on pose-free 3D reconstruction

This section evaluates previous pose-free surface reconstruction methods on reflective and textureless objects.

**Baselines.** We select DUSt3R [25], SPARF [22], Nope-NeRF [1], and CF-3DGS [8] as baselines for pose-free 3D reconstruction. Our experiments show that SPARF [22] and Nope-NeRF [1] are sensitive to pose initialization. Therefore, we initialize their poses with calibrated values while allowing them to be learned during optimization. In



Figure 5. Qualitative comparison with SuperNormal [3] (abbreviated as "SN") using surface normal maps estimated by SDM-UniPS [11] and StableNormal [27] (abbreviated as "SDM" and "ST"), respectively. The top row visualizes the input surface normals and their angular error distributions.

Table 2. Quantitative evaluation of shape and pose recovery using normal maps from SDM-UniPS [11] and StableNormal [27]. The mean angular errors (MAE) of the input normal maps are shown in the header.

Method	Metric	StableNormal [27] (MAE: 23.6°)	SDM-UniPS [11] (MAE: 8.3°)	GT Normal (MAE: 0°)
SuperNormal [3]	$\mathrm{CD}\downarrow$	0.543	0.194	0.142
PMNI		0.602	0.252	0.153
SuperNormal [3]	F1-score ↑	0.644	0.962	0.993
PMNI		0.620	0.946	0.990
PMNI	$\begin{array}{c} \text{RPEr}(^{\circ}) \downarrow \\ \text{RPEt} \downarrow \end{array}$	1.375 0.384	0.304 0.095	0.176 0.051

contrast, our method initializes camera poses in a circular distribution, as detailed in the supplementary material. DUSt3R [25] and CF-3DGS [8] do not require pose initialization.

Since SPARF [22], Nope-NeRF [1], and CF-3DGS [8] focus on pose-free novel view synthesis and output per-view depth for geometric estimation, we evaluate them quantitatively using depth maps. However, depth maps  $\hat{z}$  from both existing methods and ours have global scale ambiguity compared to the GT. We use the depth map  $z_s$  from SuperNormal [3] with calibrated camera poses as the GT reference and compute a global scale *s* that minimizes the difference between  $s\hat{z}$  and  $z_s$  using least squares. The relative depth error is then defined as the mean absolute difference between the scaled depth  $s\hat{z}$  and  $z_s$  divided by  $z_s$ .



Figure 6. Summary of RT3D dataset for pose-free reflective surface reconstruction.

Table 3. Quantitative comparison between existing methods and ours on camera pose and surface shape estimation.

Mathad			RI	PEr(°)↓			
method	MONKEY	CAT	PINEAPPLE	Dog	DRAGON	TIGER	Avg
DUSt3R [25]	3.175	2.049	2.640	2.216	2.602	4.839	2.920
Nope-NeRF [1]	9.371	8.472	7.513	8.674	8.467	8.282	8.463
SPARF [22]	7.233	6.395	3.485	3.620	<u>0.731</u>	0.695	3.693
CF-3DGS [8]	16.867	16.664	17.276	14.789	15.625	16.659	16.313
PMNI	0.230	0.356	0.258	0.258	0.439	0.582	0.354
Mathad			I	RPEt ↓			
Method	MONKEY	CAT	PINEAPPLE	Dog	DRAGON	TIGER	Avg
DUSt3R [25]	0.329	0.199	0.247	0.490	0.224	0.335	0.304
Nope-NeRF [1]	0.695	0.596	0.610	0.774	0.654	0.637	0.661
SPARF [22]	0.375	0.203	0.146	0.261	0.041	0.058	0.181
CF-3DGS [8]	0.947	0.796	1.092	0.878	0.998	1.124	0.972
PMNI	0.015	0.020	0.016	0.019	0.027	0.035	0.022
Method	Relative Depth Error $\downarrow$						
	MONKEY	CAT	PINEAPPLE	Dog	DRAGON	TIGER	Avg
DUSt3R [25]	0.062	0.056	0.046	0.147	0.046	0.075	0.072
Nope-NeRF [1]	0.276	0.191	0.305	0.489	0.231	0.176	0.278
SPARF [22]	0.099	0.055	0.038	0.131	0.029	0.050	0.067
CF-3DGS [8]	0.363	0.360	0.475	0.488	0.477	0.502	0.444
PMNI	0.011	0.017	0.008	0.010	0.011	0.026	0.014

**RT3D dataset.** To quantitatively evaluate reconstruction quality on reflective surfaces, we construct a multi-view dataset with ground-truth meshes. Fig. 6 shows our captured 6 objects with highly reflective surfaces. For each object, we use a Canon EOS R5 camera to capture 20 views surrounding the object. For each view, we take 11 images under varying illumination by moving an area light source to different positions. These multi-light images are used for photometric stereo to generate reliable surface normals.

To facilitate camera pose calibration, we place each target object on an OLED screen displaying ArUco markers, as shown in Fig. 6. The scene is captured twice, once with the display on and once off. The images with ArUco markers are used for evaluation only. Images without ArUco markers serve as input for baseline methods and our approach. Additionally, we scan the shape of the 6 objects with an EinScan SP scanner<sup>1</sup>, which provides a reference for qualitatively assessing the reconstructed shapes.

Pose evaluation. As shown in Fig. 7, we visualize the GT (shown in red) and estimated poses (shown in blue) from existing methods and ours on MONKEY and DOG object. The red line connecting the GT and estimated camera positions illustrates the performance of pose recovery. CF-3DGS [8] and Nope-NeRF [1] cannot produce reasonable pose estimation, possibly due to the temporal continuity assumption, which is not satisfied in the pose distribution of RT3D. SPARF [22] applies a pre-trained dense correspondence network, which may not generalize well on reflective and textureless surfaces such as MONKEY and DOG, affecting the pose estimation. DUSt3R [25] has relatively better results based on learned point cloud correspondence but there still remains a gap between its poses and GT. Given a circular pose initialization shown in the second column, the estimated poses from our method are accurately aligned with the corresponding GT, as shown in the last column.

As shown in the top and middle rows of Table 3, our recovered poses, including rotation and translation, achieve state-of-the-art performance over existing methods, demonstrating the strength of using multi-view surface normals for optimizing the camera poses.

**Shape evaluation.** As shown in Fig. 8, we compare estimated shapes from existing methods and ours, where DUSt3R [25] and our method can output multi-view mesh, and the shape visualizations from other methods are based on depth. Consistent with the pose estimation, DUSt3R [25] obtains better results than existing pose-free methods, but is still unsatisfactory compared with scanned meshes. In contrast, our PMNI gets detailed shape recoveries for the two reflective and textureless surfaces, and the results are close to SuperNormal [3] and scanned meshes, showing the strength of our method. More results on RT3D can be found in the supplementary material.

# 4.3. Ablation study

We conduct an ablation study to test the effectiveness of different loss terms, taking POT2 in DiLiGenT-MV [16] as an example. As shown in Table 4,  $\mathcal{L}_{normal}$  contributes most to

<sup>&</sup>lt;sup>1</sup>https://www.einscan.com/einscan-sp. Retrieved Nov. 14th, 2024.



Figure 7. Qualitative comparison of camera pose recovery on MONKEY and DOG object of RT3D dataset. The red line segment connects the calibrated and estimated camera locations to illustrate the quality of pose recovery.



Figure 8. Qualitative evaluation of shape recovery on MONKEY and DOG objects of the RT3D dataset.

Table 4. Ablation study on different loss terms.

Mathad	Shape	estimation	Poss estimation		
Method	$CD\downarrow$	F1-score $\uparrow$	$\text{RPEt}\downarrow$	$RPEr(^{\circ})\downarrow$	
Ours w/o $\mathcal{L}_{normal}$	0.691	0.769	0.139	0.542	
Ours w/o $\mathcal{L}_{ni}$	0.163	0.988	0.082	0.280	
Ours w/o $\mathcal{L}_c$	0.126	0.998	0.052	0.210	
Ours	0.115	0.999	0.037	0.141	

the shape and pose recovery, demonstrating the necessity of supervising surface normal for pose-free 3D reconstruction. Without  $\mathcal{L}_{ni}$ , the prior of integrated depth is missing, leading to an error increase in shape and pose estimation.  $\mathcal{L}_c$  based on multi-view surface normal consistency also helps to improve the accuracy. Combining these loss terms, our method get accurate shape and pose estimation.

# 5. Conclusion

We introduce PMNI, the first method that recovers both shape and pose solely from multi-view normal maps. Due

to the scarcity of features in reflective and textureless objects in the RGB domain, existing joint optimization-based methods struggle with pose and shape recovery. In contrast to RGB images, PMNI utilizes surface normals as input, which are robust to reflective and textureless surfaces. By incorporating depth from normal integration as a prior and leveraging multi-view geometric consistency, we jointly optimize shape and camera poses using a neural SDF network. We hope our method will contribute to detailed 3D reconstruction in casual capture settings.

# Acknowledgment

This work was supported by Hebei Natural Science Foundation Project No. F2024502017, Beijing-Tianjin-Hebei Basic Research Funding Program No. 242Q0101Z, National Natural Science Foundation of China (Grant No. 62472044, U24B20155, 62225601, U23B2052), Beijing Natural Science Foundation (Grant No. QY24202). We thank openbayes.com for the support of computational resources.

### References

- [1] Wenjing Bian, Zirui Wang, Kejie Li, Jia-Wang Bian, and Victor Adrian Prisacariu. Nope-nerf: Optimising neural radiance field with no pose prior. In *Proc. of IEEE Conference* on Computer Vision and Pattern Recognition (CVPR), pages 4160–4169, 2023. 1, 3, 5, 6, 7, 8
- [2] Baptiste Brument, Robin Bruneau, Yvain Quéau, Jean Mélou, François Lauze, Jean-Denis Durou, and Lilian Calvet. Rnb-neus: Reflectance and normal based reconstruction with neus. In Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2024. 1, 3
- [3] 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, 3, 5, 6, 7, 8
- [4] Xu Cao, Hiroaki Santo, Boxin Shi, Fumio Okura, and Yasuyuki Matsushita. Bilateral normal integration. In Proc. of European Conference on Computer Vision (ECCV), 2022. 2, 4
- [5] Guangcheng Chen, Yicheng He, Li He, and Hong Zhang. Pisr: Polarimetric neural implicit surface reconstruction for textureless and specular objects. *Proc. of European Conference on Computer Vision (ECCV)*, 2024. 3
- [6] Pinxuan Dai, Jiamin Xu, Wenxiang Xie, Xinguo Liu, Huamin Wang, and Weiwei Xu. High-quality surface reconstruction using gaussian surfels. In ACM SIGGRAPH 2024 Conference Papers. Association for Computing Machinery, 2024. 2
- [7] Akshat Dave, Yongyi Zhao, and Ashok Veeraraghavan. Pandora: Polarization-aided neural decomposition of radiance. In *Proc. of European Conference on Computer Vision* (ECCV), pages 538–556. Springer, 2022. 3
- [8] Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A. Efros, and Xiaolong Wang. Colmap-free 3d gaussian splatting. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20796–20805, 2024. 1, 3, 6, 7, 8
- [9] 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. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4251–4260, 2023. 1, 2
- [10] Yufei Han, Heng Guo, Koki Fukai, Hiroaki Santo, Boxin Shi, Fumio Okura, Zhanyu Ma, and Yunpeng Jia. Nersp: Neural 3d reconstruction for reflective objects with sparse polarized images. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11821–11830, 2024. 1, 3
- [11] Satoshi Ikehata. Scalable, detailed and mask-free universal photometric stereo. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13198– 13207, 2023. 2, 3, 6
- [12] Kaiwen Jiang, Yang Fu, Mukund Varma T, Yash Belhe, Xiaolong Wang, Hao Su, and Ravi Ramamoorthi. A constructoptimize approach to sparse view synthesis without camera pose. In *Proc. of SIGGRAPH*, pages 1–11, 2024. 3

- [13] Yakun Ju, Ling Li, Xian Zhong, Yuan Rao, Yanru Liu, Junyu Dong, and Alex C. Kot. Underwater surface normal reconstruction via cross-grained photometric stereo transformer. *IEEE Journal of Oceanic Engineering*, 50(1):192– 203, 2025. 1
- [14] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 9492–9502, 2024. 3
- [15] Chenhao Li, Taishi Ono, Takeshi Uemori, Hajime Mihara, Alexander Gatto, Hajime Nagahara, and Yusuke Moriuchi. Neisf: Neural incident stokes field for geometry and material estimation. In *Proc. of IEEE Conference on Computer Vision* and Pattern Recognition (CVPR), pages 21434–21445, 2024.
- [16] 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. 5, 6, 7
- [17] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In Proceedings of the IEEE/CVF international conference on computer vision, pages 5741–5751, 2021. 1, 3
- [18] 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. ACM Trans. on Graph., 42(4):1–22, 2023. 1, 2
- [19] 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 Proc. of European Conference on Computer Vision (ECCV), 2020. 2
- [20] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 12179–12188, 2021. 1, 3
- [21] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 3
- [22] Prune Truong, Marie-Julie Rakotosaona, Fabian Manhardt, and Federico Tombari. Sparf: Neural radiance fields from sparse and noisy poses. IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2023. 3, 6, 7, 8
- [23] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions* on Pattern Analysis and Machine Intelligence, 13(04):376– 380, 1991. 5
- [24] 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. *Proc. of Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 1, 2, 3
- [25] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d

vision made easy. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20697–20709, 2024. 1, 2, 3, 6, 7, 8

- [26] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2024. 1
- [27] Chongjie Ye, Lingteng Qiu, Xiaodong Gu, Qi Zuo, Yushuang Wu, Zilong Dong, Liefeng Bo, Yuliang Xiu, and Xiaoguang Han. Stablenormal: Reducing diffusion variance for stable and sharp normal. ACM Trans. on Graph., 2024. 6