GCRayDiffusion: Pose-Free Surface Reconstruction via Geometric Consistent Ray Diffusion

Li-Heng Chen[1,2] Zi-Xin Zou[2] Chang Liu[1] Tianjiao Jing[1] Yan-Pei Cao[2] Shi-Sheng Huang[1]†
Hongbo Fu[3] Hua Huang[1]

[1]Beijing Normal University [2]VAST [3]Hong Kong University of Science and Technology

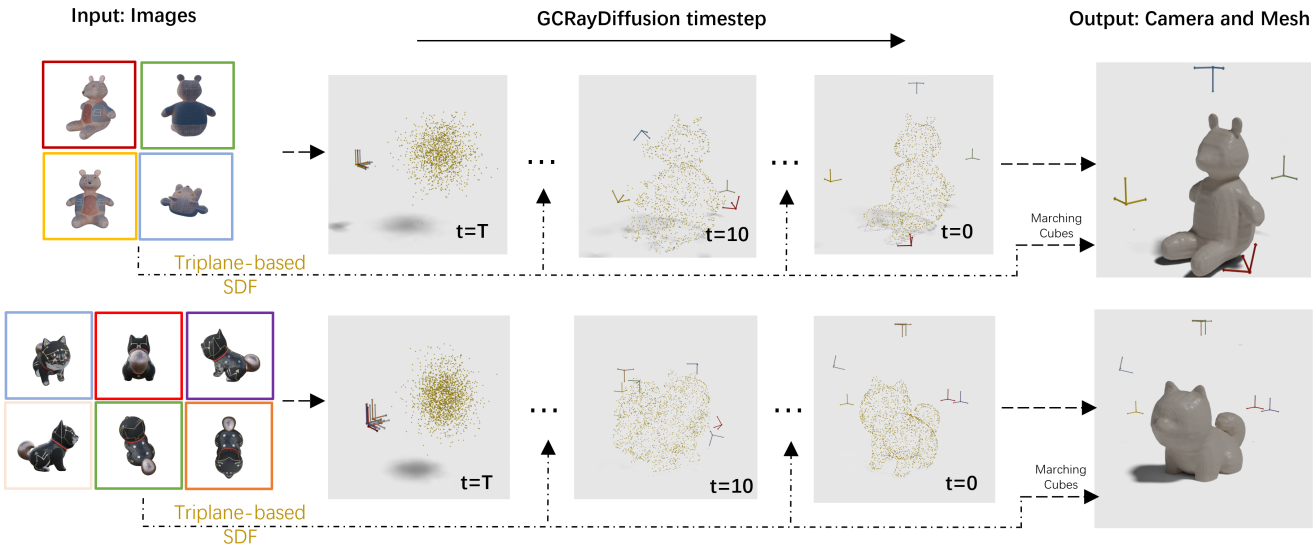

Figure 1. We achieve accurate pose-free neural surface learning with the aid of a novel geometric consistent ray diffusion, i.e., GCRayDiffusion, even from sparse view images (left column). Our GCRayDiffusion model formulates the images' camera poses as neural ray bundles and provides *explicit* sampling points generated during the denoiser processing (middle columns) to regularize the triplane-based SDF learning, achieving accurate surface reconstruction and camera pose estimation simultaneously (right column).

Abstract

Accurate surface reconstruction from unposed images is crucial for efficient 3D object or scene creation . However, it remains challenging particularly for the joint *camera pose estimation. Previous approaches have achieved impressive pose-free surface reconstruction results in dense-view settings but could easily fail for sparse-view scenarios without sufficient visual overlap. In this paper, we propose a new technique for pose-free surface reconstruction, which follows triplane-based signed distance field (SDF) learning but regularizes the learning by* explicit *points sampled from ray-based diffusion of camera pose estimation. Our key contribution is a novel Geometric Consistent Ray Diffusion model (GCRayDiffusion), where we represent camera poses as neural bundle rays and regress the distribution of noisy rays via a diffusion model. More importantly, we further condition the denoising process of RGRayDiffusion using the triplane-based SDF of the entire scene, which provides effective 3D consistent regularization to get multi-view consistent camera pose estimation. Finally, we incorporate RGRayDiffusion to the triplane-based SDF learning by introducing on-surface geometric regularization from the sampling points of the neural bundle rays, which leads to highly accurate pose-free surface reconstruction results even for sparse view inputs. Extensive evaluations on public datasets show that our GCRayDiffusion achieves more accurate camera pose estimation than previous approaches,*

† Corresponding authors.

with geometrically more consistent surface reconstruction results, especially given sparse view inputs.

1. Introduction

3D surface reconstruction from multi-view images has been a long-standing research topic in computer graphics and vision communities. It serves as a crucial 3D content creation tool for various applications such as VR/AR, video games, and robotics. We have seen significant progress made from the recent neural surface reconstruction using deep implicit representation [1, 18, 35, 36], neural radiance field (NeRF) [20, 33, 46–48, 61], and 3D Gaussian Splatting (3DGS) [21–23, 31, 57]. However, most of these approaches rely on highly accurate camera pose information as input for each image and would easily fail given less accurate camera poses like noisy views or unknown camera poses.

For pose-free surface reconstruction, one traditional solution is to first estimate the camera poses using the Structure-of-Motion (S*f*M) technique [40, 44] and then perform surface reconstruction according to the estimated camera poses. However, the vanilla S*f*M technique needs dense viewpoints between images with sufficient overlap and would cause unsatisfactory pose estimation for sparse view images with little visual overlap. For robust pose estimation from sparse view images, subsequent works directly regress the camera pose parameters from wide baseline images [2, 4, 8, 38, 58], predict the relative pose probability distributions [6, 25, 65], or use an iterative refinement strategy [43], but the pose estimation quality is still limited. Recent works represent the camera poses as a joint distribution conditioned on image observations [52] or rays [66] and regress the camera poses via the denoiser process of diffusion models, achieving impressive camera pose estimation results. However, such diffusion-aided approaches still rely on dense feature matching [52] and fail to perform effective bundle adjustment for sparse view scenarios [66].

On the other hand, some recent works propose to jointly learn the neural surface representation and camera poses, leveraging the geometric cues from photometric [7, 26, 32, 56, 62], silhouettes [3, 24, 64], or depth points [55]. However, those joint learning strategies are only performed independently across dense input images. Some subsequent works [16, 19, 26, 32, 51, 54] further explore the extra relations across multiple views to optimize both the neural surface representation and camera poses, achieving more accurate neural surface reconstruction results. However, these approaches still need accurate geometric priors, such as depth [51] within sufficient overlaps [16] or extra camera intrinsic information [54]. They could not guarantee geometrically consistent surface reconstruction quality given sparse view inputs in many highly freeform applications with unbounded scenarios.

We propose a new pose-free surface reconstruction approach, which leverages effective diffusion-based bundle adjustment to achieve multi-view consistent camera pose estimation and simultaneously leads to geometric consistent surface reconstruction quality even given sparse view inputs. Based on a triplane-based SDF learning of an entire scene from multiple images, we incorporate geometric priors from multi-view consistent camera pose bundle adjustment to regularize the neural implicit field learning. Inspired by the recent ray-based camera parametrization [66], we introduce a new neural bundle ray representation to over-parameterize camera poses but with an extra depth attribution. Leveraging the depth information, we can trace the end points of the neural bundle rays, which can serve as the *explicit* sampling points of the on-surface geometry, thus enabling a differentiable connection between the camera pose and neural implicit field representation. More importantly, we build a Geometric Consistent Ray Diffusion (GCRayDiffusion) model to regress the noisy rays, conditioned on the triplane-based SDF of the entire scene for multi-view consistent camera pose estimation. Finally, we incorporate the denoiser process of GCRayDiffusion to the triplane-based SDF learning by leveraging the on-surface geometry regularization from the sampling points of the neural bundle rays. Our approach leads to multi-view consistent camera pose estimation and geometric consistent surface reconstruction simultaneously, as shown in Fig. 1.

To evaluate the effectiveness, we perform extensive evaluations of our approach on publicly released datasets, such as the Objaverse dataset [10] and the Google Scanned Object (GSO) [12] dataset, by comparing with state-of-the-art camera pose estimation approaches, including COLMAP [40], RelPose++ [25], PoseDiffusion [52], RayDiffusion [66] and neural surface reconstruction approaches, such as FORGE [19], DUSt3R [55]. According to the quantitative and qualitative comparisons, our approach achieves much better robustness and accuracy in camera pose estimation than those previous approaches, and also geometrically more consistent surface reconstruction results, especially given sparse view image inputs.

2. Related Work

2.1. Camera Pose Estimation

The classical Structure-from-Motion (S*f*M) [40, 44] has been a traditional solution to estimate camera poses from unordered images, which basically relies on finding feature points [30] in overlapping images and performs camera poses optimization using Bundle Adjustment [50]. Subsequent approaches have significantly improve the SfM quality by improving the feature quality [11], correspondences [39, 42, 60] and differentiable Bundle Adjust-

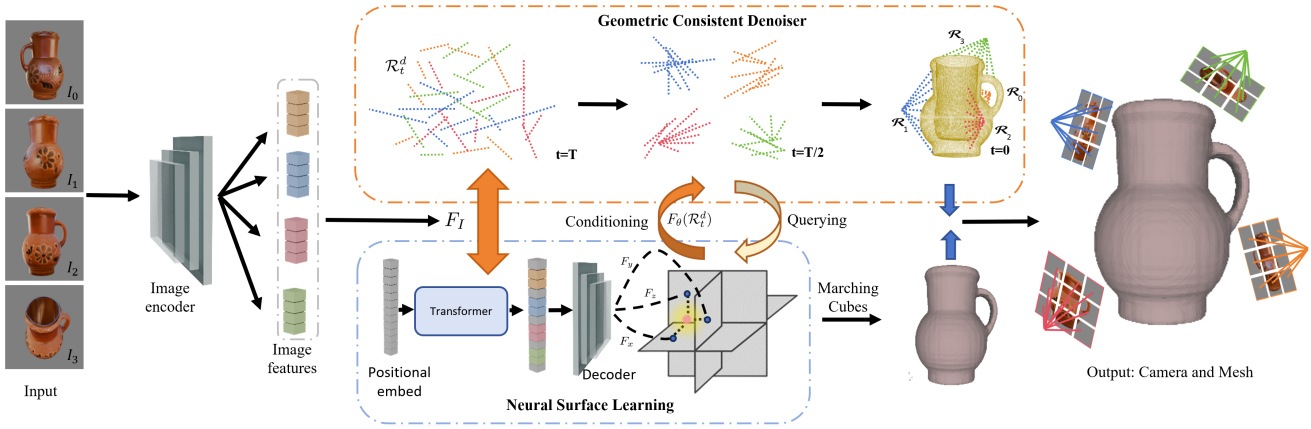

Figure 2. The pipeline of our GCRayDiffusion. Given sparse view images $\mathcal{I}$, our approach extract the image features $F_{\mathcal{I}}$ using an image encoder, and feed $F_{\mathcal{I}}$ to two sub-branches: (1) Geometric Consistent Denoiser processing, which regresses the neural ray bundles $\mathcal{R}_t^d$ following a SDF conditioned ray-based diffusion, to estimate the camera poses, and (2) Neural Surface Learning of a triplane-based SDF $F_\theta(\mathcal{R}_t^d)$. During the ray bundles denoiser processing, we generate explicit sampling points from neural ray bundles to regularizing the neural surface learning, by querying their SDFs from the triplane-based SDF and locating their position on the surface of the object shape, which leads to accurate surface reconstruction and camera poses estimation simuntaneously.

ment [27, 49]. However, the SfM framework still rely on dense feature points to estimate camera poses for images with sufficient overlaps, which would lead to significant quality decrease for sparse view images with little overlaps.

To perform camera poses estimation from sparse view images, recent efforts have explored to directly regress 6DoF camera poses from sparse images [2, 4, 8, 38, 58], or predict the probabilistic distribution of relative pose [25, 65] using energy-based models. SparsePose [43] proposed to iteratively refine the sparse camera poses from the initial estimation. RelPose++ [25] further defines a new camera pose coordinate system and decouples the rotation and translation prediction for more robust camera pose estimation. More recently, with the success of Diffusion models [15, 45], PoseDiffusion [52] proposed to regress the camera pose using a diffusion-aided bundle adjustment. Zhang et al [66] introduce bundle rays for even sparse view image inputs. Our approach for camera pose estimation is inspired by these previous approaches, but leverage the geometric prior guidance to the ray based diffusion to achieve multi-view consistent camera pose estimation.

2.2. Neural Surface Reconstruction

There have already been significant progress made for neural surface reconstruction from image sets, by representing scene geometry as deep implicit representation [1, 18, 35, 36], NeRF [33, 47, 53, 61] or 3D Gaussian Splatting [21, 23, 31, 63]. Subsequent works further incorporate more explicit surface supervisions [13], surface rendering [34] or multi-view geometry priors [9] for more accurate surface learning. However, most of these previous works dense input views for accurate neural surface learning, which would not work for sparse view inputs scenarios. Recently, SparseNeuS [28] achieves more generalizable neural surface learning form sparse input views, but still relies on highly accurate camera poses. Unlike these previous neural surface reconstruction works, our approach enables geometric consistent surface reconstruction directly from unposed sparse images, which performs camera pose estimation using a ray based diffusion during the neural surface learning simultaneously.

2.3. Joint Implicit Learning and Pose Optimization

Another category approaches for pose-free surface reconstruction is to jointly perform implicit field learning and camera pose optimization. BARF [26] would be probably one of the first works to the adjust the camera pose directly on NeRF representation following a coarse-to-fine registration strategy. GARF [7] further improve the robustness of camera pose refinement using Gaussian based activation functions. SCNeRF [17] proposed to optimize the ray intersection re-projection error during the NeRF learning to adjust camera poses, with subsequent efforts made for more accurate joint learning leveraging more geometric cues such as silhouette [3, 24] or semantic mask [64]. However, most of the approaches depends on dense input views [59] to perform the joint implicit learning and pose estimation, which will not be effective for sparse scenarios [5].

Recently, for sparse view scenarios, SPARF [51] proposed to jointly learn the neural surface and refine camera poses using the depth priors. SC-NeuS [16] introduced a joint learning of camera poses and deep implicit representation via the explicit regularization from on-surface geometry. FORGE [19] established cross-view correlations to es-

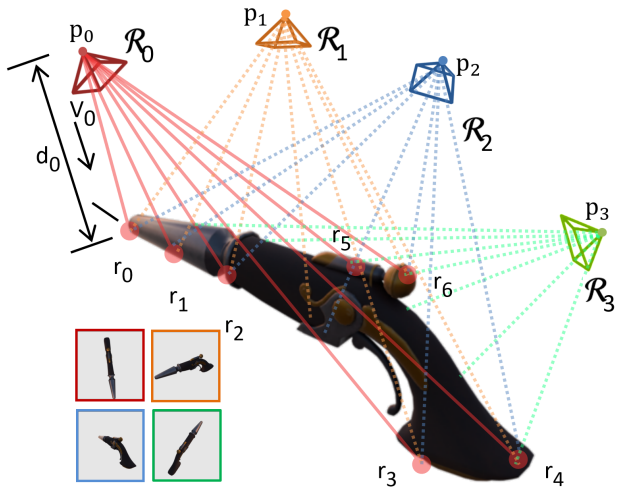

Figure 3. The illustration of our neural bundle rays definition.

timate relative camera poses, which in turn improves the object surface learning. PF-LRM [54] and DUSt3R [55] predict sparse poses by predicting pixel-aligned pointclouds and using PnP to recover cameras.

Inspired by those previous work, our approach proposes to incorporate the explicit regularization from diffusion-based camera pose estimation, to the triplane-based signed distance field learning, which achieves more geometric consistent surface reconstruction results with multi-view consistent camera pose estimation at the same time.

3. Method

Given a sparsely sampled image set $\mathcal{I} = \{I_i | i = 1, ..., N\}$ with N is the image number, our goal is to estimate the camera poses $\mathcal{T} = \{T_i | i = 1, ..., N\}$ for each image I_i, and perform the surface reconstruction represented as neural signed distance field (SDF), i.e., $F_\theta(x \in R^3 | \theta) = s \in R$ with θ represents the parameters of the deep network and s is the signed distance value.

Inspired by previous approach [14, 41, 66], we first propose a new camera pose parametrization using a set of neural bundle rays $\mathcal{R} = \{\mathbf{r}_k | k = 1, ..., M\}$ to parameterize each camera, with the key difference that each ray $\mathbf{r}_k$ is additionally affiliated with depth information (Section 3.1). Secondly, we formulate the distribution of noisy rays conditioned on image feature embedding and the signed distance field F_θ, and model the camera pose estimation process as a Geometric Consistent Ray Diffusion model (GCRayDiffusion), to recover camera poses $\mathcal{T}$ by learning the denoiser process of the GCRayDiffusion (Section 3.2). Finally, we construct the triplane-based SDF learning by cooperating the on-surface geometry regularization from the sampling points of the neural bundle rays, which can lead to multi-view consistent camera pose estimation and geometric consistent neural surface learning simultaneously (Section 3.3). Fig. 2 illustrates the main pipeline of our approach.

3.1. Neural Bundle Ray Representation

Unlike most of the previous approaches that representing camera poses T_i with a 6DoF vector (including 3D rotation and translation), we follow the latest ray-based parametrization introduced by [66] for a more flexible camera pose representation. But different from [66], we additionally record the depth information, which indicates the distance from the intersect point that the ray intersect with the shape surface. In this way, we can *explicitly* trace the on-surface sampling points for each ray, thus constructing a differentiable bridge between camera pose and surface representation for thereafter diffusion-aided neural surface learning.

Specifically, as shown in Fig. 3, we propose to over-parameterize each image I_i as a set of neural bundle rays $\mathcal{R}_i = \{\mathbf{r}_k^i | k = 1, ..., M\}$, with each ray $\mathbf{r}_k^i$ is represented as a 7-dimension vector including a unit directional vector $\mathbf{v}_k^i \in R^3$ though any point $\mathbf{p}_k^i \in R^3$ and depth $d_k^i \in R$ following Plücker coordinates [37] as:

$$\mathbf{r}_k^i = (\mathbf{v}_k^i, \mathbf{m}_k^i, \mathbf{d}_k^i) \in R^7, \tag{1}$$

where $\mathbf{m}_k^i = \mathbf{p}_k^i \times \mathbf{v}_k^i \in R^3$ is the moment vector. Given an image I_i with known camera pose, we can uniformly sample a set of 2D pixel coordinates $\{u_k\}_M$ to construct the neural bundle rays $\mathcal{R}_i$, and compute the unit directional vector $\mathbf{v}_i^k \in R^3$ by unprojecting rays from the pixel coordinates, where the moment vectors $\mathbf{m}_k^i$ can be computed by treating the camera centers as the point p since all rays intersect at the camera center. Conversely, given a collection of neural bundle rays $\mathcal{R}_i$ associated with 2D pixels $\{u_k\}_M$, we can recover the camera extrinsic and intrinsic by solving the intersection of all rays in $\mathcal{R}_i$. Please refer to our supplementary materials for the detailed derivations.

Important Property. Another important difference of our neural bundle ray representation from previous approaches [66] is that we can trace the end point $\mathbf{r}_d \in R^3$ for each ray $\mathbf{r}$, since we record an additional depth information d. Specifically, we can compute each ray's end point as $\mathbf{r}_d = d \cdot \mathbf{v} + \mathbf{p}$. The end point $\mathbf{r}_d \in R^3$ can also been seen as the intersection point that $\mathbf{r}$ intersects with the object's surface, thus connecting the camera pose and object's surface differentiablly, which serves an important property for thereafter ray diffusion aided neural surface learning.

3.2. Geometric Consistent Ray Diffusion

Based on our above neural bundle ray representation, we view the bundle adjustment process of the noisy rays during the camera pose estimation as a diffusion process, and recover the final rays from initial noisy rays by reversing the Markovian noising process. Specifically, for noisy ray distribution $\mathcal{R}_t \sim q(\mathcal{R}_t), t = 0, ..., T$, the noising stepsize

in the diffusion process is defined by a variance schedule $\{\beta_t\}_{t=0}^T$ as:

$$q(\mathcal{R}_t|\mathcal{R}_{t-1}) = \mathcal{N}(\mathcal{R}_t; \sqrt{(1-\beta_t)}\mathcal{R}_{t-1}, \beta_t\mathbf{I}), \quad (2)$$

where $q(\mathcal{R}_t|\mathcal{R}_{t-1})$ is a normal distribution, such that the noisy rays can be computed as:

$$\mathcal{R}_t = \sqrt{\bar{\alpha}_t}\mathcal{R}_0 + \epsilon\sqrt{1-\bar{\alpha}_t}, \quad (3)$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = 1 - \prod_{s=0}^{t} \alpha_s$, and the noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$.

Geometric Consistent Denoiser. Unlike previous diffusion-based camera pose estimation approaches [52, 66] which directly learn a vanilla denoiser of the target data distribution, we propose to learn a geometric consistent denoiser, which predicts the noise from noisy rays conditioned on the signed distance field (SDF) F_θ of the entire shape. By leveraging such globally consistent geometry prior to the denoiser process, the ray distribution can be effectively bundle-adjusted, thus yielding to multi-view consistent rays prediction for high accurate camera pose estimation. Specifically, as shown in Fig. 2, we learn a denoiser network g_ϕ to predict the noise ϵ added in the most recent rays $\mathcal{R}_t$ as:

$$g_\phi(\mathcal{R}_t, t|F_\theta(\mathcal{R}_t^d), F_I) \rightarrow \epsilon, \quad (4)$$

where $\mathcal{R}_t^d$ is the sampling points set of the rays set $\mathcal{R}_t$, i.e., $\mathcal{R}_t^d = \{\mathbf{r}_d\}$ with each sampling point $\mathbf{r}_d$ is computed from the corresponding ray $\mathbf{r}^j \in \mathcal{R}_t$, $F_\theta(\mathcal{R}_t^d)$ is the signed distance value for the sampling points predicted by F_θ, F_I is the feature vector extracted from the original image I at the 2D-pixel coordinate of each ray. To train the denoiser network g_ϕ, we utilize the L2 distance as loss during the parameters optimization following:

$$L_{diff} = ||g_\phi(\mathcal{R}_t, t|F_\theta(\mathcal{R}_t^d), F_I) - \epsilon||_2.$$

In this way, we build up a geometric consistent Ray Diffusion (GCRayDiffusion) model to predict the neural bundle rays set of each input image, and recover the corresponding camera pose via the transformation of the neural bundle rays.

3.3. Diffusion-aided Neural Surface Learning

Finally, we incorporate the GCRayDiffusion model the the neural surface learning for the surface reconstruction. Our key observation is to leverage the sampling points of the neural bundle rays as explicit regularization to guide the neural signed distance field (SDF) F_θ learning, thus introducing the multi-view consistent camera pose bundle adjustment priors via GCRayDiffusion for the F_θ learning, towards geometric consistent surface reconstruction results.

Specifically, as shown in Fig. 2, we formulate F_θ as a triplane-based signed distance field (SDF), which consists a Transformer-based image encoder Φ to extract triplane feature maps from image inputs and a MLP-based decoder $\mathcal{D}$ to regress the SDF prediction, as:

$$F_\theta(x \in R^3|\Phi, \mathcal{D}) \rightarrow s \in R, \quad (5)$$

$$s.t. \quad \Phi(\mathcal{I}) = \{F_x, F_y, F_z\}, \quad \mathcal{D}(x \in R^3|F_x, F_y, F_z) = s,$$

where $\{F_x, F_y, F_z\}$ are the triplane feature maps and s is the SDF value.

Diffusion-aided Learning. We leverage the sampling points $\mathcal{R}^d$ of the neural bundle rays $\mathcal{R}$ during the T step denoiser process of the GCRayDiffusion model, to guide $F_\theta(x \in R^3|\Phi, \mathcal{D})$ learning. One straightforward yet effective operation is to regularize such sampling points $\mathcal{R}_t^d$ located on the latent geometry surface of $F_\theta(x \in R^3|\Phi, \mathcal{D})$, i.e.,

$$F_\theta(\mathcal{R}_t^d|\Phi, \mathcal{D}) \rightarrow 0, \quad (6)$$

during each time step t of the denoiser process from the GCRayDiffusion model. So after T step denoising process, we effectively guide the neural SDF learning using the multi-view consistent camera pose bundle adjustment, which leads to more better geometric consistent surface learning. Once $F_\theta(\mathcal{R}_t^d|\Phi, \mathcal{D})$ is learned, we extract the surface results of the zero level-set of $F_\theta(\mathcal{R}_t^d|\Phi, \mathcal{D})$ using Marching Cubes [29]. Simultaneously, the final neural bundle rays $\mathcal{R}$ are recovered by the GCRayDiffusion model, which is used to compute the final camera poses $\mathcal{T}$.

4. Experiments

4.1. Experimental Setup

Dataset and Metrics. We evaluate our approach on the Objaverse dataset [10] and the Google Scanned Object (GSO) [12] dataset. The Objaverse dataset consists of diverse 3D scenes, while the GSO dataset includes 300 samples from unseen, allowing us to test the generalization ability of our method. To evaluate the accuracy for camera pose estimation, we adopt two accuracy metrics including camera rotation accuracy (within 15 degrees), camera translation accuracy (within 0.1). For the surface reconstruction accuracy, we adopt the following metrics including Hausdorff Distance (HD), Chamfer Distance (CD), Normal Consistency (NC), by measuring the surface mesh extracted from our GCRayDiffusion and the ground truth surface mesh, and also F-score (based on the Hausdorff Distance accuracy) where we use HD threshold as 5% when calculating the Precision and Recall respectively.

Training Details. Our training process involves initializing camera poses, for the Objaverse dataset, we utilize

random initialization. The training consists of 40K iterations, with 25K allocated to the coarse stage and 15K to the fine stage. The weights for the loss functions are set as follows: $w_A = 0.1$, $w_\rho = w_M = 0.001$, and $\lambda_{np} = 0.05$. The local weights are $\lambda_{ml} = \lambda_{nl} = 0.01$, and the flatten weight is $\lambda_f = 20$. All experiments are conducted on a single NVIDIA A800 GPU.

Comparing Approaches. We compare our method with state-of-the-art camera pose estimation approaches, including RelPose++ [25], PoseDiffusion [52], RayDiffusion [66], FORGE [19] and DUSt3R [55]. Besides, we also choose COLMAP [40] as baseline approach during the evaluation. During the experiment, we evaluate both the camera pose estimation and surface reconstruction accuracy. For the camera pose, we directly compute the accuracy metrics (both rotation and translation) for these comparing approaches. As for the surface reconstruction, since COLMAP [40], RelPose++ [25], PoseDiffusion [52] and RayDiffusion [66] only compute camera poses but didn't perform surface reconstruction. For a fair comparison, we further conduct NeRF-based surface reconstruction using their camera poses estimation. Then we perform the mesh surface quality extracted from all of these comparing approaches to conduct the comparison. We use the public release source code of COLMAP, RelPose++, PoseDiffusion, RayDiffusion, FORGE and DUSt3R by using the default parameter configuration for fair comparison. To achieve better performance for COLMAP, we use SuperPoint features [11] and SuperGlue matching [39] during the experiments.

4.2. Evaluation on Objaverse Dataset

We first conduct evaluation on Objaverse dataset, by comparing our approaches with those previous approaches. Since the cases in Objaverse dataset have different number of input images, for a comprehensive evaluation, we conduct experiments by changing the number of input images, i.e., from 2-6 images, for all of the different comparing approaches.

Camera Pose Estimation Comparison. As shown in Table 1, in terms of camera rotation accuracy and camera translation accuracy, our approach can achieve consistently better accuracy than all of those previous approaches, where our approach significantly outperforms COLMAP, RelPose++ and FORGE respectively, and also better camera pose estimation in rotation and translation accuracy than SOTA approaches such as PoseDiffusion, RayDiffusion and DUSt3R respectively. Besides, given different number of image input (from 2 to 6), our approach also consistently outperform those previous approaches.

	Rotation Accuracy				
# of images	2	3	4	5	6
COLMAP	31.20	30.16	28.74	29.89	30.69
RelPose++	61.35	62.71	65.79	66.11	68.42
FORGE	89.26	89.89	88.36	79.23	78.65
PoseDiffusion	77.3	74.82	75.25	69.34	62.1
RayDiffusion	86.00	85.00	87.20	80.33	79.39
DUSt3R	90.52	91.87	92.26	91.58	91.16
Ours	**93.21**	**93.17**	**92.32**	**93.60**	**92.92**
	Translation Accuracy				
# of images	2	3	4	5	6
COLMAP	29.36	26.25	21.83	23.79	25.21
RelPose++	63.24	60.55	57.31	58.12	57.46
FORGE	48.54	44.39	41.33	43.58	43.26
PoseDiffusion	40.21	39.33	38.23	31.25	30.88
RayDiffusion	65.32	50.43	41.28	39.91	39.80
DUSt3R	68.26	62.45	62.03	62.97	60.21
Ours	**69.77**	**63.44**	**62.62**	**63.91**	**62.89**

Table 1. **The camera pose estimation accuracy evaluated on Objaverse dataset.**

Surface Reconstruction Comparison. Except from the camera pose estimation, we also conduct comparison on the surface reconstruction accuracy. As shown in Table 3 (upper rows), our approach also achieves consistently better accuracy metrics, including CD, HD, NC and F-scores, than all of those previous approaches.

Qualitative Comparison. The qualitative results presented in Fig. 4 illustrate the high-quality surface reconstructions achieved by our method from Objaverse dataset, and also some of those previous SOTA approaches such as RelPose++ (with NeRF reconstruction), FORGE and DUSt3D respectively. As we can see in the figure, RelPose++ often crush to achieve a complete surface reconstruction, though the camera poses estimation are reasonable, but NeRF fails to conduct success surface reconstruction given such sparse image input. Although FORGE and DUSt3R can achieve reasonable surface reconstruction results, but our approach can achieve accurate surface reconstruction with more geometric details, with the benefit of more accurate camera pose estimation.

4.3. Generalization Evaluation to GSO Dataset

We also evaluate the generalization ability of our approach to another GSO dataset, where we use the parameter weights pre-trained on Objaverse dataset and conduct test on GSO dataset. Besides, we also make comparison with those previous approaches mentioned above.

https://github.com/colmap/colmap
https://github.com/amyxlase/relpose-plus-plus
https://github.com/facebookresearch/PoseDiffusion
https://github.com/jasonyzhang/RayDiffusion
https://github.com/UT-Austin-RPL/FORGE
https://github.com/naver/dust3r

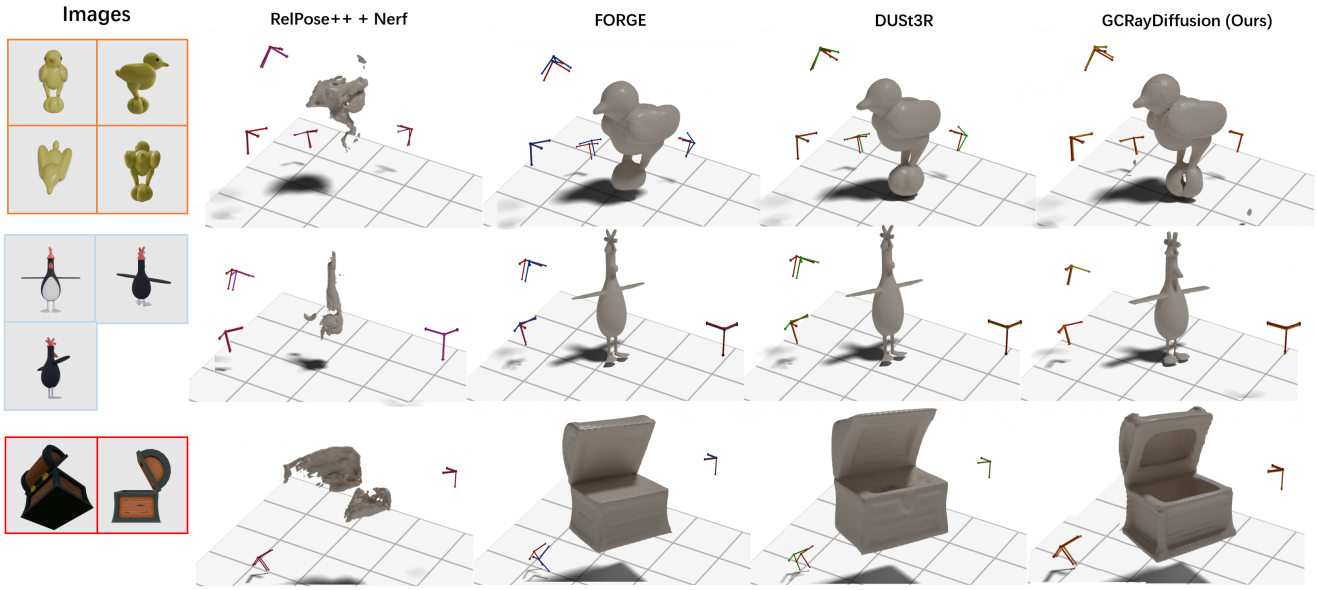

Figure 4. Qualitative surface reconstruction comparison evaluated on Objaverse dataset for different comparing approaches, including RelPose++, FORGE, DUSt3R and our GCRayDiffusion (from left to right column) respectively.

Camera Pose Estimation Comparison. As shown in Table 2, in terms of camera rotation accuracy and camera translation accuracy, our approach can also achieve much better accuracy than all of those previous approaches, where we only achieve slightly worse camera rotation accuracy than DUSt3R when given 4 input images. This means that our approach also achieve consistently better than all of those previous approaches evaluated on GSO dataset.

Surface Reconstruction Comparison. Except from the camera pose estimation, we also conduct comparison on the surface reconstruction accuracy. As shown in Table 3 (bottom rows), our approach also achieves consistently better accuracy metrics, including CD, HD, NC and F-scores, than all of those previous approaches, where we only achieve a slightly worse F-score accuracy than DUSt3R.

Qualitative Comparison. The qualitative results presented in Fig. 5 illustrate the high-quality surface reconstructions achieved by our method from the GSO dataset, and also some of those previous SOTA approaches such as RelPose++ (with NeRF reconstruction), FORGE, and DUSt3D respectively. Similarly, our approach can also achieve better visual reconstruction results than those three SOTA approaches.

# of images	**Rotation Accuracy**				
	2	3	4	5	6
COLMAP	29.23	29.58	31.45	32.15	32.50
RelPose++	59.23	59.88	62.49	64.12	66.92
FORGE	83.21	84.37	85.96	79.83	76.11
PoseDiffusion	75.48	74.99	73.31	70.08	61.25
RayDiffusion	86.33	84.89	87.31	81.22	76.3
DUSt3R	91.33	91.27	**92.37**	90.06	91.03
Ours	**93.20**	**93.23**	91.35	**91.03**	**94.32**
# of images	**Translation Accuracy**				
	2	3	4	5	6
COLMAP	26.54	23.18	21.83	22.47	20.16
RelPose++	65.33	62.29	60.36	61.25	64.31
FORGE	49.23	48.56	45.88	46.21	42.19
PoseDiffusion	41.33	38.79	39.31	34.27	29.06
RayDiffusion	63.42	50.25	41.79	38.41	38.02
DUSt3R	66.33	61.9	61.93	59.82	60.59
Ours	**68.82**	**62.77**	**61.95**	**63.13**	**64.81**

Table 2. **The camera pose estimation accuracy evaluated on the GSO dataset.**

4.4. Ablation

We designed an ablation experiment to study how the two main components impact the final camera pose estimation and surface reconstruction respectively, including (1) how the ray diffusion performs without the condition of triplane-based SDF (termed as 'w/o SDF') for the camera pose estimation, and (2) how the triplane-based SDF learning perform without the aid of our GCRayDiffusion (termed as 'w/o ray diffuser') for the surface reconstruction. As shown in Table 4 and Table 5 evaluated on the Objaverse dataset, we can see that both the camera pose estimation and surface reconstruction quality will decrease without using the two main components. Fig. 6 also show several surface reconstruction results with or without using the guide of our GCRayDiffusion respectively.

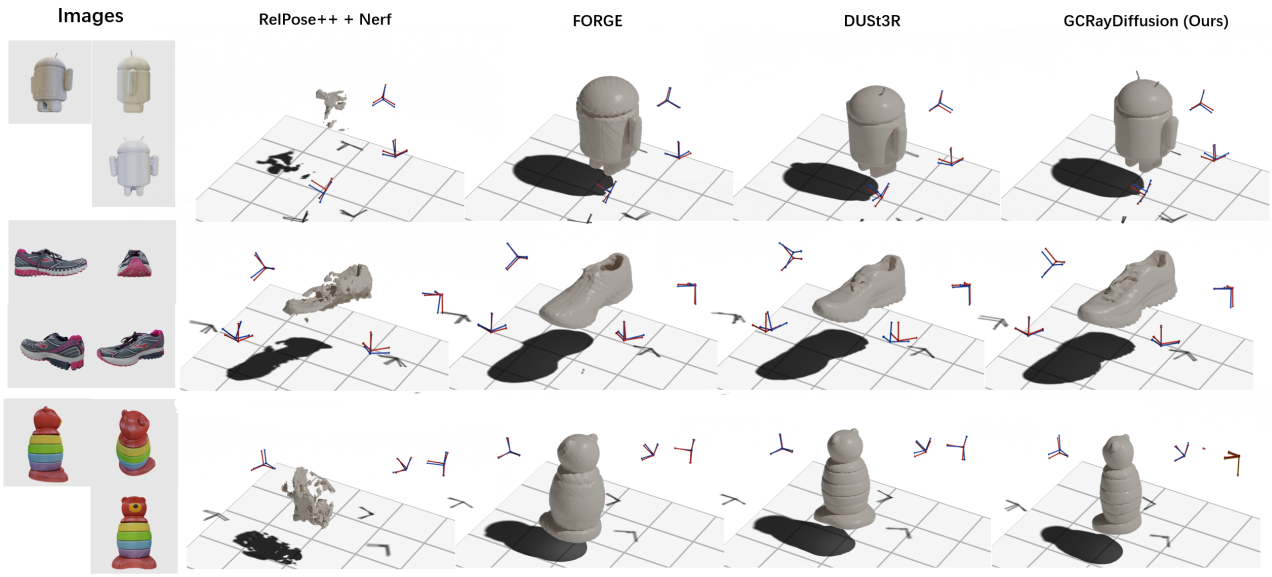

Figure 5. Qualitative surface reconstruction comparison evaluated on GSO dataset for different comparing approaches, including RelPose++, FORGE, DUSt3R, and our GCRayDiffusion (from left to right column) respectively.

Dataset	CD↓	HD↓	NC↑	F-score↑
Objaverse				
COLMAP	9.17	15.41	0.74	0.536
RelPose++	4.58	6.49	0.76	0.615
FORGE	0.145	0.405	0.989	0.89
DUSt3R	0.132	0.368	0.995	0.97
Ours	**0.125**	**0.323**	**0.997**	**0.99**
GSO				
COLMAP	10.26	13.32	0.72	0.519
RelPose++	3.96	5.72	0.73	0.527
FORGE	0.155	0.437	0.985	0.854
DUSt3R	0.139	0.366	0.973	**0.961**
Ours	**0.131**	**0.302**	**0.988**	0.958

Table 3. **Surface reconstruction accuracy on Objaverse and GSO dataset respectively.**

	Rotation	Trans
w/o SDF	86.3	37.5
GCRaydiffusion	92.32	62.62

Table 4. **Camera pose estimation accuracy comparison.**

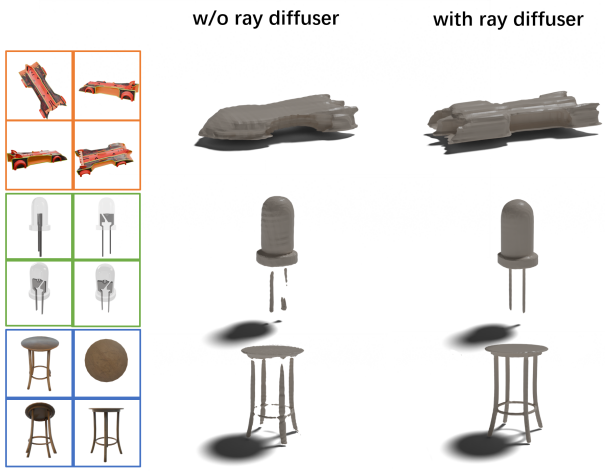

Figure 6. Surface reconstruction comparison with or without using ray diffuser of our GCRayDiffusion.

	CD↓	HD↓	NC↑	F-score↑
w/o ray diffuser	0.16	0.802	0.992	0.988
GCRaydiffusion	0.125	0.323	0.997	0.99

Table 5. **Surface reconstruction accuracy comparison.**

5. Conclusion

This paper contributes a new pose-free surface learning with the aid of a novel geometric consistent ray diffusion, i.e., GCRayDiffusion, which achieves better camera pose estimation and surface reconstruction than previous SOTA approaches. We hope that our approach can inspire subsequent works for more robust and accurate pose-free surface reconstruction from sparse image inputs in this community.

References

[1] Matan Atzmon and Yaron Lipman. Sal: Sign agnostic learning of shapes from raw data. In *IEEE CVPR*, pages 2565–2574, 2020. 2, 3

[2] Vassileios Balntas, Shuda Li, and Victor Prisacariu. Relocnet: Continuous metric learning relocalisation using neural nets. In *ECCV*, pages 751–767, 2018. 2, 3

[3] Mark Boss, Andreas Engelhardt, Abhishek Kar, Yuanzhen Li, Deqing Sun, Jonathan T Barron, Hendrik Lensch, and Varun Jampani. Samurai: Shape and material from unconstrained real-world arbitrary image collections. In *Advances in Neural Information Processing Systems*, 2022. 2, 3

[4] Ruojin Cai, Bharath Hariharan, Noah Snavely, and Hadar Averbuch-Elor. Extreme rotation estimation using dense correlation volumes. In *IEEE CVPR*, pages 14566–14575, 2021. 2, 3

[5] 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 *IEEE CVPR*, pages 14124–14133, 2021. 3

[6] Kefan Chen, Noah Snavely, and Ameesh Makadia. Wide-baseline relative camera pose estimation with directional learning. In *IEEE CVPR*, pages 3258–3268, 2021. 2

[7] Shin-Fang Chng, Sameera Ramasinghe, Jamie Sherrah, and Simon Lucey. Gaussian activated neural radiance fields for high fidelity reconstruction and pose estimation. In *ECCV*, pages 264–280. Springer, 2022. 2, 3

[8] Sungjoon Choi, Qian-Yi Zhou, and Vladlen Koltun. Robust reconstruction of indoor scenes. In *IEEE CVPR*, pages 5556–5565, 2015. 2, 3

[9] François Darmon, Bénédicte Bascle, Jean-Clément Devaux, Pascal Monasse, and Mathieu Aubry. Improving neural implicit surfaces geometry with patch warping. In *IEEE CVPR*, pages 6260–6269, 2022. 3

[10] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *IEEE CVPR*, pages 13142–13153, 2023. 2, 5

[11] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *IEEE CVPR Workshop*, pages 224–236, 2018. 2, 6

[12] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022. 2, 5

[13] Qiancheng Fu, Qingshan Xu, Yew Soon Ong, and Wenbing Tao. Geo-neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction. *Advances in Neural Information Processing Systems*, 35:3403–3416, 2022. 3

[14] Michael D Grossberg and Shree K Nayar. A general imaging model and a method for finding its parameters. In *IEEE ICCV*, pages 108–115, 2001. 4

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. Red Hook, NY, USA, 2020. Curran Associates Inc. 3

[16] Shi-Sheng Huang, Zixin Zou, Yichi Zhang, Yan-Pei Cao, and Ying Shan. Sc-neus: Consistent neural surface reconstruction from sparse and noisy views. In *AAAI*, pages 2357–2365, 2024. 2, 3

[17] Yoonwoo Jeong, Seokjun Ahn, Christopher Choy, Anima Anandkumar, Minsu Cho, and Jaesik Park. Self-calibrating neural radiance fields. In *IEEE CVPR*, pages 5846–5854, 2021. 3

[18] Chiyu Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, Thomas Funkhouser, et al. Local implicit grid representations for 3d scenes. In *CVPR*, pages 6001–6010, 2020. 2, 3

[19] Hanwen Jiang, Zhenyu Jiang, Kristen Grauman, and Yuke Zhu. Few-view object reconstruction with unknown categories and camera poses. In *IEEE 3DV*, pages 31–41, 2024. 2, 3, 6

[20] Kaiwen Jiang, Shu-Yu Chen, Feng-Lin Liu, Hongbo Fu, and Lin Gao. Nerffaceediting: Disentangled face editing in neural radiance fields. In *ACM SIGGRAPH Asia*, pages 1–9, 2022. 2

[21] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 42(4), 2023. 2, 3

[22] Bernhard Kerbl, Andreas Meuleman, Georgios Kopanas, Michael Wimmer, Alexandre Lanvin, and George Drettakis. A hierarchical 3d gaussian representation for real-time rendering of very large datasets. *ACM TOG*, 43(4):1–15, 2024.

[23] Tobias Kirschstein, Simon Giebenhain, Jiapeng Tang, Markos Georgopoulos, and Matthias Nießner. Gghead: Fast and generalizable 3d gaussian heads. In *ACM SIGGRAPH Asia*, pages 1–11, 2024. 2, 3

[24] Zhengfei Kuang, Kyle Olszewski, Menglei Chai, Zeng Huang, Panos Achlioptas, and Sergey Tulyakov. Neroic: neural rendering of objects from online image collections. *ACM TOG*, 41(4):1–12, 2022. 2, 3

[25] Amy Lin, Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose++: Recovering 6d poses from sparse-view observations. In *IEEE 3DV*, pages 106–115, 2024. 2, 3, 6

[26] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *IEEE CVPR*, pages 5741–5751, 2021. 2, 3

[27] Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. Pixel-perfect structure-from-motion with featuremetric refinement. In *IEEE ICCV*, pages 5987–5997, 2021. 3

[28] 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. Springer, 2022. 3

[29] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Sem-*

inal graphics: pioneering efforts that shaped the field, pages 347–353. 1998. 5

[30] Bruce D Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI*, pages 674–679, 1981. 2

[31] Xiaoyang Lyu, Yang-Tian Sun, Yi-Hua Huang, Xiuzhe Wu, Ziyi Yang, Yilun Chen, Jiangmiao Pang, and Xiaojuan Qi. 3dgsr: Implicit surface reconstruction with 3d gaussian splatting. *ACM TOG*, 43(6):1–12, 2024. 2, 3

[32] Quan Meng, Anpei Chen, Haimin Luo, Minye Wu, Hao Su, Lan Xu, Xuming He, and Jingyi Yu. Gnerf: Gan-based neural radiance field without posed camera. In *IEEE ICCV*, pages 6351–6361, 2021. 2

[33] 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. *Communications of the ACM*, 65(1):99–106, 2021. 2, 3

[34] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *IEEE ICCV*, pages 5589–5599, 2021. 3

[35] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *IEEE CVPR*, pages 165–174, 2019. 2, 3

[36] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *ECCV*, pages 523–540. Springer, 2020. 2, 3

[37] Julius Plücker. *Analytisch-Geometrische Entwicklungen*. GD Baedeker, 1828. 4

[38] Chris Rockwell, Justin Johnson, and David F Fouhey. The 8-point algorithm as an inductive bias for relative pose prediction by vits. In *IEEE 3DV*, pages 1–11, 2022. 2, 3

[39] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *IEEE CVPR*, pages 4938–4947, 2020. 2, 6

[40] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *IEEE CVPR*, pages 4104–4113, 2016. 2, 6

[41] Thomas Schops, Viktor Larsson, Marc Pollefeys, and Torsten Sattler. Why having 10,000 parameters in your camera model is better than twelve. In *IEEE CVPR*, pages 2535–2544, 2020. 4

[42] Xi Shen, François Darmon, Alexei A Efros, and Mathieu Aubry. Ransac-flow: generic two-stage image alignment. In *ECCV*, pages 618–637. Springer, 2020. 2

[43] Samarth Sinha, Jason Y Zhang, Andrea Tagliasacchi, Igor Gilitschenski, and David B Lindell. Sparsepose: Sparse-view camera pose regression and refinement. In *IEEE CVPR*, pages 21349–21359, 2023. 2, 3

[44] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM SIGGRAPH*, pages 835–846. 2006. 2

[45] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, pages 2256–2265. PMLR, 2015. 3

[46] Nagabhushan Somraj and Rajiv Soundararajan. Vip-nerf: Visibility prior for sparse input neural radiance fields. In *ACM SIGGRAPH*, pages 1–11, 2023. 2

[47] Jiaming Sun, Xi Chen, Qianqian Wang, Zhengqi Li, Hadar Averbuch-Elor, Xiaowei Zhou, and Noah Snavely. Neural 3d reconstruction in the wild. In *ACM SIGGRAPH*, pages 1–9, 2022. 3

[48] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, et al. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH*, pages 1–12, 2023. 2

[49] Chengzhou Tang and Ping Tan. Ba-net: Dense bundle adjustment network. *arXiv preprint arXiv:1806.04807*, 2018. 3

[50] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. In *Vision Algorithms: Theory and Practice: International Workshop on Vision Algorithms Corfu, Greece, September 21–22, 1999 Proceedings*, pages 298–372. Springer, 2000. 2

[51] Prune Truong, Marie-Julie Rakotosaona, Fabian Manhardt, and Federico Tombari. Sparf: Neural radiance fields from sparse and noisy poses. In *IEEE CVPR*, pages 4190–4200, 2023. 2, 3

[52] Jianyuan Wang, Christian Rupprecht, and David Novotny. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *IEEE ICCV*, pages 9773–9783, 2023. 2, 3, 5, 6

[53] 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. *Advances in Neural Information Processing Systems*, 34:27171–27183, 2021. 3

[54] Peng Wang, Hao Tan, Sai Bi, Yinghao Xu, Fujun Luan, Kalyan Sunkavalli, Wenping Wang, Zexiang Xu, and Kai Zhang. Pf-lrm: Pose-free large reconstruction model for joint pose and shape prediction. In *ICLR*, 2024. 2, 4

[55] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *IEEE CVPR*, pages 20697–20709, 2024. 2, 4, 6

[56] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. NeRF−−: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. 2

[57] Tong Wu, Yu-Jie Yuan, Ling-Xiao Zhang, Jie Yang, Yan-Pei Cao, Ling-Qi Yan, and Lin Gao. Recent advances in 3d gaussian splatting. *Computational Visual Media*, 10(4):613–642, 2024. 2

[58] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *Robotics: Science and Systems XIV*, 2018. 2, 3

[59] Yuxi Xiao, Nan Xue, Tianfu Wu, and Gui-Song Xia. Level-sfm: Structure from motion on neural level set of implicit surfaces. In *IEEE CVPR*, pages 17205–17214, 2023. 3

[60] Gengshan Yang and Deva Ramanan. Volumetric correspondence networks for optical flow. *Advances in neural information processing systems*, 32, 2019. 2

[61] Guo-Wei Yang, Wen-Yang Zhou, Hao-Yang Peng, Dun Liang, Tai-Jiang Mu, and Shi-Min Hu. Recursive-nerf: An efficient and dynamically growing nerf. *IEEE TVCG*, 29(12): 5124–5136, 2022. 2, 3

[62] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33:2492–2502, 2020. 2

[63] Zehao Yu, Torsten Sattler, and Andreas Geiger. Gaussian opacity fields: Efficient adaptive surface reconstruction in unbounded scenes. *ACM TOG*, 43(6):1–13, 2024. 3

[64] Jason Zhang, Gengshan Yang, Shubham Tulsiani, and Deva Ramanan. Ners: neural reflectance surfaces for sparse-view 3d reconstruction in the wild. *Advances in Neural Information Processing Systems*, 34:29835–29847, 2021. 2, 3

[65] Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose: Predicting probabilistic relative rotation for single objects in the wild. In *ECCV*, pages 592–611. Springer, 2022. 2, 3

[66] Jason Y Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras as rays: Pose estimation via ray diffusion. In *ICLR*, 2024. 2, 3, 4, 5, 6