SC-NeuS: Consistent Neural Surface Reconstruction from Sparse and Noisy Views

Shi-Sheng Huang[1], Zixin Zou[2], Yichi Zhang[3], Yan-Pei Cao[4], Ying Shan[4]
[1]Beijing Normal University
[2]Tsinghua University
[3]Beijing Institute of Technology
[4] ARC Lab, Tencent PCG
huangss@bnu.edu.cn, zouzx19@mails.tsinghua.edu.cn, zhangyc@bit.edu.cn, caoyanpei@gmail.com, yingsshan@tencent.com

Abstract

The recent neural surface reconstruction approaches using volume rendering have made much progress by achieving impressive surface reconstruction quality, but are still limited to dense and highly accurate posed views. To overcome such drawbacks, this paper pays special attention on the consistent surface reconstruction from sparse views with noisy camera poses. Unlike previous approaches, the key difference of this paper is to exploit the multi-view constraints directly from the explicit geometry of the neural surface, which can be used as effective regularization to jointly learn the neural surface and refine the camera poses. To build effective multi-view constraints, we introduce a fast differentiable on-surface intersection to generate on-surface points, and propose view-consistent losses on such differentiable points to regularize the neural surface learning. Based on this point, we propose a joint learning strategy, named SC-NeuS, to perform geometry-consistent surface reconstruction in an end-to-end manner. With extensive evaluation on public datasets, our SC-NeuS can achieve consistently better surface reconstruction results with fine-grained details than previous approaches, especially from sparse and noisy camera views. The source code is available at https://github.com/zouzx/sc-neus.git.

Introduction

3D surface reconstruction from multi-view images continues to be an important research topic in computer vision and graphics communities. Unlike traditional Multi-View Stereo (MVS) based methods leveraging structure from motion (S*f*M) (Snavely, Seitz, and Szeliski 2006) technique for sparse (Labatut, Pons, and Keriven 2007; Schönberger et al. 2016; Schonberger and Frahm 2016; Xu and Tao 2019) or dense (Kar, Häne, and Malik 2017; Yao et al. 2018; Xu and Tao 2020) surface reconstruction, the recent neural surface reconstruction approaches (Yariv et al. 2020; Wang et al. 2021a; Oechsle, Peng, and Geiger 2021; Huang et al. 2021; Azinović et al. 2022; Darmon et al. 2022; Fu et al. 2022; Zou et al. 2022) adopt to learn the deep implicit representation (Park et al. 2019; Peng et al. 2020; Atzmon and Lipman 2020; Jiang et al. 2020) with the aid of volume rendering (Mildenhall et al. 2021), leading to more better *complete* and *fine-grained* surface reconstruction quality, which have received much research attention for multi-view image based 3D reconstruction.

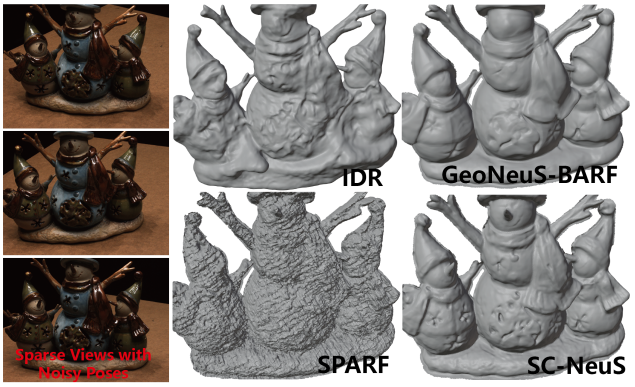

Figure 1: Surface reconstruction results from sparse input images with noisy camera poses. By exploring the multi-view constraints directly from the explicit geometry of neural surface, our SC-NeuS can achieve much better surface reconstruction than previous approaches like IDR (Yariv et al. 2020), GeoNeuS (Fu et al. 2022) and SPARF (Truong et al. 2023) etc. Note that GeoNeuS with the coarse-to-fine learning of BARF (Lin et al. 2021) (termed as GeoNeuS-BARF) also fails to achieve satisfactory results compared to our SC-NeuS.

As like NeRF (Mildenhall et al. 2021), one main drawback of most neural surface reconstruction approaches (NeuS (Wang et al. 2021a), VolSDF (Yariv et al. 2021), Unisurf (Oechsle, Peng, and Geiger 2021), NeuralWarp (Darmon et al. 2022), GeoNeuS (Fu et al. 2022)) is the dependency on dense input views, which is not suitable for many real-world applications with only sparse input views and often noisy camera poses. Some subsequent works propose to improve the reconstruction quality from sparse scenarios, by introducing regularization like sparse points (Deng et al. 2022), multi-views depth priors (Chen et al. 2021; Niemeyer et al. 2022; Truong et al. 2023), rendering ray entropy (Kim, Seo, and Han 2022) or geometry-aware feature volume (Long et al. 2022). However, most of these approaches are still relying on highly accurate camera poses, which could not be easily obtained using technique

like COLMAP (2016) for sparse input views.

To overcome the dependency on highly accurate cameras poses, many recent works propose to jointly learn the deep implicit geometry and the camera poses, with the guidance of novel registration from photometric (Chng et al. 2022; Lin et al. 2021; Meng et al. 2021; Wang et al. 2021c; Yariv et al. 2020) or silhouette (Boss et al. 2022; Kuang et al. 2022; Zhang et al. 2021) priors. But since those registrations are often performed *independently* across dense input views, the registration quality would significantly drop for sparse scenarios (Fig. 1), where enough *relations* across views are missing to effectively bundle adjust both the deep implicit geometry and camera poses. It still remains to be challenging to jointly learn the deep implicit geometry and camera poses from sparse input views (Zhang, Ramanan, and Tulsiani 2022) for geometry-consistent surface reconstruction.

This paper proposes a **S**parse-view **C**onsisent **Neu**ral **S**urface (SC-NeuS) learning strategy, which performs geometry-consistent surface reconstruction with fine-grained details from sparse and noisy camera poses (as few as 3 views). Unlike previous *independent* registrations from dense input views, we seek to explore more effective multi-view constraints between sparse views. Due to the gap between the volume rendering integral and point-based SDF modeling (Fu et al. 2022), except from relying on the depth constraints (Chen et al. 2021) rendered from the under-constrained signed distance field (Fu et al. 2022), we utilize extra regularization directly from the explicit geometry of the neural surface representation. Our key insight is that the observation of the explicit surface geometry across multiple views should be *consistent*, which can be used as effective regularization to jointly learn both the neural surface representation and camera poses. Specifically, we first introduce a fast *differentiable* on-surface intersection to sample on-surface points from explicit geometry of the neural surface, and then provide effective view-consistent losses defined on such differentiable on-surface intersections, which builds up an end-to-end joint learning for the neural surface representation and camera poses. Besides, to further improve the geometry-consistent neural surface learning, we incorporate an coarse-to-fine learning strategy (Lin et al. 2021) for highly accurate and fine-grained surface reconstruction results.

To evaluate the effectiveness of our SC-NeuS, we conduct extensive experiments on public dataset including DTU (Jensen et al. 2014) and BlendedMVS (Yao et al. 2020) with various geometry scenarios. Compared with previous state-of-the-art approaches(Lin et al. 2021; Fu et al. 2022; Jeong et al. 2021; Yariv et al. 2020; Wang et al. 2021a), our SC-NeuS achieves consistently better geometry-consistent surface reconstruction results both quantitatively and qualitatively, which becomes a new state-of-the-art neural surface reconstruction approach from sparse and noisy cameras.

Related Work

Novel View Synthesis. The success of NeRF (Mildenhall et al. 2021) and subsequent works (Trevithick and Yang 2021; Wang et al. 2021b; Yu et al. 2021) have achieved impressive novel view synthesis applications. To overcome the drawback of dense input views, multiple works propose to extra regularizations or priors for sparse view novel view synthesis, such as depth and appearance smoothness (RegNeRF (Niemeyer et al. 2022), MVSNeRF (Chen et al. 2021)), ray entropy regularization (InfoNeRF (Kim, Seo, and Han 2022)), perceptual losses (SVS (González et al. 2022)), Spatio-Temporal consistency (Li et al. 2023) or ray distortion (Mip-NeRf360 (Barron et al. 2022)) et al. Besides, some recent approaches (Wei et al. 2021; Deng et al. 2022; Roessle et al. 2022) use depth priors to constrain the NeRF optimization, which also achieves promising novel view synthesis results from sparse input views. Different from these previous approaches, our approaches aims at geometry-consistent neural surface learning with noisy camera poses, and contributes a joint neural surface learning and camera pose optimization strategy from sparse input views.

Neural Surface Reconstruction. The neural implicit representation (DeepSDF (Park et al. 2019; Jiang et al. 2020)) or NeRF (Mildenhall et al. 2021) has been state-of-the-art way for neural surface reconstruction. IDR (Yariv et al. 2020) introduces a neural surface rendering based on the neural implicit representation, which enables precise surface learning from 2D images. The NeRF based approaches (NeuS (Wang et al. 2021a)) further incorporate more explicit surface supervisions (GeoNeuS (Fu et al. 2022)), balance between surface rendering and volume rendering (UNISURF (Oechsle, Peng, and Geiger 2021)) or multi-view geometry priors (NeuralWarp (Darmon et al. 2022)) for more accurate surface learning. However, most of these previous works depend on dense input views for accurate neural surface learning, which is not feasible for sparse scenarios.

Recently, SparseNeuS (Long et al. 2022) achieves more generalizable neural surface learning form sparse input views, but still relies on highly accurate camera poses. In contrast, our approaches enables accurate neural surface learning from sparse input views, and optimizes the noisy camera poses simultaneously.

Joint Deep Implicit Geometry and Pose Optimization. BARF (Lin et al. 2021) is probably one of the first works to reduce NeRF's dependent on highly accurate camera poses, by introducing a coarse-to-fine registration strategy. GARF (Chng et al. 2022) further provides a Gaussian based activation functions for more robust camera pose refinement. SCNeRF (Jeong et al. 2021) builds geometric loss optimization on the ray intersection re-projection error. Subsequent works (Boss et al. 2022; Kuang et al. 2022; Zhang et al. 2021) also incorporate the photometric loss from silhouette or mask, but requires accurate foreground segmentation. However, most of these approaches still depends on dense input views (Level-S^2fM (Xiao et al. 2023)), which will not be effective for sparse scenarios.

Different from these previous approaches, our approach explores the view-consistent constraints on the explicit surface geometry of neural surface representation, which provides more effective cues than rendered depth (Truong et al. 2023) to jointly learn neural surface and refine camera poses in an end-to-end manner, without need any shape prior (Zhang et al. 2021) or RGB-D input (Sucar et al. 2021; Azinović et al. 2022; Zhu et al. 2022).

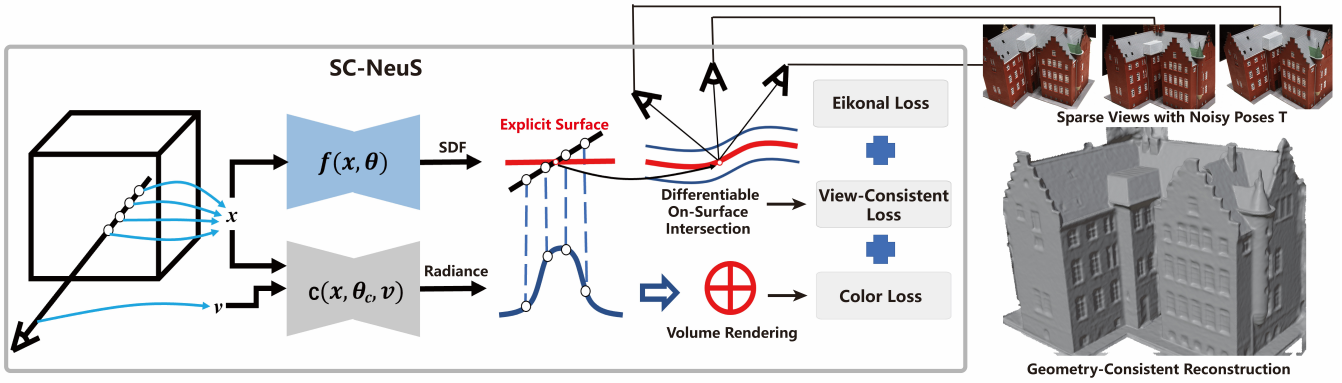

Figure 2: The overview of our SC-NeuS. Given sparse input images (as few as 3 views) with noisy camera poses T, our SC-NeuS represents the object's geometry as a signed distance field $f(x, \theta)$ and perform volume rendering of the geometry with an extra radiance field $c(x, \theta_c, v)$. Combing the effective view-consistent loss defined on the multi-view differentiable intersection points on the explicit surface of $f(x, \theta)$, with color loss and Eikonal loss, our SC-Neus performs jointly learning of $f(x, \theta)$, $c(x, \theta_c, v)$ and camera poses T in an end-to-end manner, achieving geometry-consistent surface reconstruction results with fine-grained details.

SC-NeuS

Given sparse view images (as few as 3) with noisy camera poses of an object, we aim at reconstructing the surface represented by neural implicit function and jointly optimizing the camera poses. Specifically, for sparse input views $I = \{I_i\}$ with noisy camera poses $T = \{T_i\}$ ($i \in \{1, 2, 3\}$), we adopt to represent the object's geometry as signed distance field (SDF) $f(x, \theta)$ ($x \in R^3$, θ is the MLP parameter), and render its appearance using volume rendering from an extra radiance field $c(x, \theta_c, v)$ as provided by NeuS (Wang et al. 2021a).

By introducing effective multi-view constraints across sparse views, we propose an new joint learning strategy, called SC-NeuS, for both signed distance field $f(x, \theta)$ learning and camera poses $T = \{T_i\}$ optimization. Fig. 2 demonstrates the main pipeline of our SC-NeuS framework in an end-to-end learning manner.

From Multi-view Constraints to Geometry-consistent Surface Learning. Unlike the previous approaches (Chng et al. 2022; Lin et al. 2021; Meng et al. 2021; Yariv et al. 2020) using photometric loss across dense input views *independently*, we adopt to exploit multi-view constraints as extra effective regularization to constraint the surface learning. Besides, instead of relying on multi-view depth rendering prior from the neural surface to multi-view depth priors (Chen et al. 2021; Truong et al. 2023; Fu et al. 2022), we propose to utilize more multi-view regularization directly from the explicit surface geometry of the neural surface. Our key observation is that the geometry cues (points or patches) locating on the shape surface should be consistently observed across multi-views, which is intuitively an effective constraints for geometry-consistent surface learning, especially in sparse scenarios.

Specifically, we first derive an fast *differentiable* point intersection on the explicit surface of signed distance field $f(x, \theta)$. Then we provide view-consistent losses for two kinds of on-surface geometry cues (3D sparse points and patches) based on our differentiable point intersection, including view-consistent re-projection loss and patch-warping loss , to effectively regularize the joint learning of signed distance field $f(x, \theta)$ and camera poses T . Since the intersection derived by our approach is differentiable for both the neural surface parameters θ and camera poses T, our neural surface learning can be performed in an end-to-end manner without any other supervisions.

Differentiable On-surface Intersection

To enable multi-view consistent constraints, the essential requirement of the geometry cues is that they need *locate* on the explicit surface, i.e., the zero level set of the signed distance field $f(x, \theta)$. Considering a 2D feature point $p \in R^2$ in the reference image I_i with camera pose T_i, we seek to compute its intersection point $P^* \in R^3$ on the surface geometry of signed distance field $f(x, \theta)$. According to volume rendering of SDF (Mildenhall et al. 2021; Wang et al. 2021a), there exists a ray length value t^* such that:

$$P^* = c_i + t^* v, \quad f(P^*, \theta) = 0,$$

where c_i and v are the camera center point and casting ray of p respectively.

Although IDR (Yariv et al. 2020) have provided a differentiable intersection derivation for P^*, however, which is somewhat too slow to enable an efficient neural surface learning. Therefore, we propose a new differentiable on-surface intersection for fast neural surface learning. Specifically, as shown in Fig. 3, we first uniformly sample points in the casting ray v of 2D feature point p with sampling depth value set $\mathbf{T} = \{t_k\}$. Then we find the depth value t_k such that $f(c_i + t_k v, \theta) f(c_i + t_{k+1} v, \theta) < 0$. Finally, we move t_k along the casting ray v to the on-surface intersection $P^*(T_i, \theta, v)$ following:

$$P^*(T_i, \theta, v) = c_i + t_k v - \frac{v}{\langle \frac{\partial f}{\partial x}, v \rangle} f(c_i + t_k v, \theta). \quad (1)$$

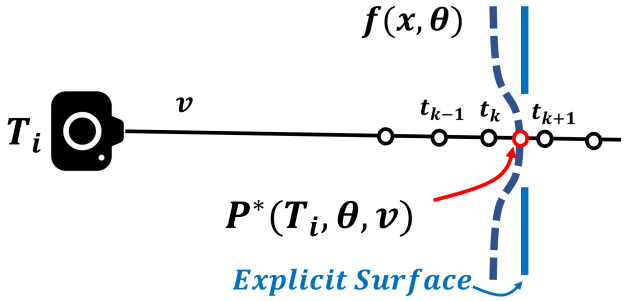

Figure 3: The illustration of our fast differentiable on-surface intersection $P^*(T_i, \theta, v)$, between the explicit surface of signed distance field $f(x, \theta)$ and camera view T_i along the casting ray v.

View-Consistent Loss

Based on our differentiable intersection, we further define effective losses to neural surface learning in the multi-view scenario. Specifically, we utilize two kinds of on-surface geometry cues, i.e., 3D sparse points and patches (Fig. 4), and formulate view-consistent losses for these on-surface geometry cues respectively.

View-consistent Re-projection Loss. Considering a pair of 2D feature correspondence (p_k^i, p_k^j) from reference image I_i (camera pose T_i) and target image I_j (camera pose T_j) with $p_k^i \in I_i, p_k^j \in I_j$, we compute the on-surface intersection 3D point P_k^{ij} via our differentiable intersection. By re-projecting P_k^{ij} back to I_i and I_j, we get the re-projection location as $\bar{p}_k^i = \pi(P_k^{ij}, T_i)$, $\bar{p}_k^j = \pi(P_k^{ij}, T_j)$, where $\pi(\cdot)$ is the camera projection operator. For a geometry-consistent surface reconstruction, the re-projection error between $p_k^i \to \bar{p}_k^i$ and $p_k^j \to \bar{p}_k^j$ should be minimized. Then we formulate the view-consistent re-projection loss L_r for all of possible sparse correspondence as:

$$L_r = \sum_{i,j} \sum_{k \in N_k} (|p_k^i - \pi(P_k^{ij}, T_i)| + ||p_k^j - \pi(P_k^{ij}, T_j)|).$$

View-consistent Patch-warping Loss. We also consider the on-surface patch (Fig. 4) to utilize the geometric structure constraints to further improve the neural surface learning. Specifically, we warp the on-surface patch to multi-view images but in a differentiable way using our differentiable multi-view intersection. For a small patch s on the surface which is observed by image pair I_i, I_j, we represent the plane equation of s in the camera coordinate of the reference image I_i as $n^T p + d = 0$, where $p(T_i, T_j)$ is the intersection point from I_i, I_j, n is the normal computed with automatic differentiation of the signed distance field $f(x, \theta)$ at $p(T_i, T_j)$. Suppose that the s is projected to I_i, I_j to obtain image patches $s_i \in I_i, s_j \in I_j$ respectively, for image pixel $x \in s_i$ and its corresponding pixel $x' \in s_j$, we have:

$$x = Hx', H = K_i(R_i R_j^t - \frac{R_i(R_i^T t_i - R_j^T t_j)n^T}{d})K_j^{-1},$$

where H is the homography matrix, $T_i = \{R_i|t_i\}, T_j = \{R_j|t_j\}$, K_i, K_j are the intrinsic camera matrix for image pair I_i, I_j.

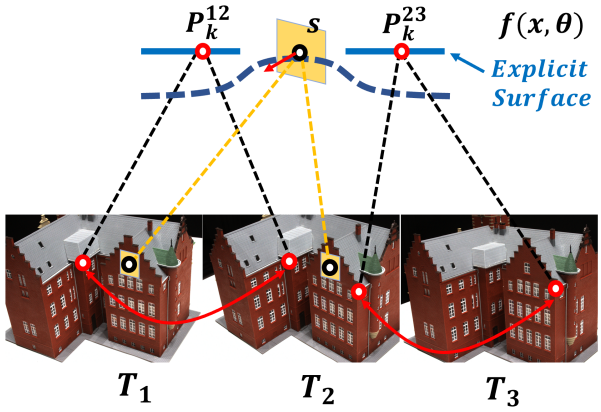

Figure 4: An example illustration of view-consistent losses defined on two kinds of on-surface geometry cues, i.e., 3D sparse points (colored red) and patches (colored orange), for a sparse view (3 views) input case from DTU dataset.

We use the normalization cross correlation (NCC) as the view-consistent patch-warping loss as :

$$L_{ncc}(s_i, s_j) = \frac{Cov(I_i(s_i), I_j(s_j))}{Var(I_i(s_i))Var(I_j(s_j))},$$

where Cov and Var donates the covariance and variance for color identity of patches (s_i, s_j) respectively.

Training Strategy

Based on the view-consistent losses, we formulate the objective function E as:

$$E = L_{color} + \lambda_r L_r + \lambda_{ncc} L_{ncc} + \lambda_{reg} L_{reg}, \quad (2)$$

with L_r and L_{ncc} are the view-consistent re-projection loss and patch-warping loss defined above, and L_{color} and L_{reg} are the color rendering loss and Eikonal regularization loss proposed by NeuS (Wang et al. 2021a) as:

$$L_{color} = \frac{1}{N} \sum_i^N |\mathcal{R}(f(x,\theta), c(x,\theta_c,v), T_i) - I_i|,$$

$$L_{reg} = \frac{1}{M} \sum_i^M (|| \bigtriangledown f_\theta||_2 - 1)^2,$$

where $\mathcal{R}(f(x,\theta), c(x,\theta_c,v), T_i)$ is the volume rendering image from $f(x,\theta), c(x,\theta_c,v)$ to view T_i.

So in summary, we propose to jointly learn the signed distance field $f(x,\theta)$, radian field $c(x,\theta_c,v)$ and camera poses $T = \{T_i\}$ to optimize the objective function E in an end-to-end manner following:

$$\{\theta^*, \theta_c^*, T^*\} = \arg\min_{\theta,\theta_c,T} E. \quad (3)$$

Please refer to our supplementary materials for more details on the network training and coarse-to-fine learning strategy.

	Translation ↓															
Scan	24	37	40	55	63	65	69	83	97	105	106	110	114	118	122	Mean
BARF	42.3	56.4	17.1	46.4	13.1	0.72	21.7	48.7	3.33	48.3	44.9	46.2	22.8	1.03	13.2	28.4
IDR	41.8	41.8	43.4	42.7	1.12	0.39	0.82	1.12	10.1	2.95	59.5	47.7	46.7	0.62	27.1	24.5
NBF	30.7	39.5	27.1	10.3	24.2	3.47	23.8	53.2	13.8	6.72	38.7	14.0	0.34	31.7	30.4	23.2
NBF*	45.3	39.9	41.7	43.9	1.16	0.48	1.28	1.76	1.73	1.44	53.3	46.9	0.17	0.66	29.6	20.6
GBF	**0.12**	52.0	18.1	29.3	**0.07**	**0.06**	0.27	52.3	**0.27**	0.08	28.5	0.20	39.2	0.20	10.5	15.4
SPF	0.24	0.29	0.44	0.09	0.32	0.61	0.21	0.24	0.57	0.40	**0.10**	0.59	0.31	0.14	**0.07**	0.31
Ours	0.15	**0.23**	**0.16**	**0.07**	0.16	0.17	**0.16**	**0.07**	0.31	**0.01**	0.17	**0.12**	**0.23**	**0.12**	0.17	**0.15**
	Rotation (°) ↓															
Scan	24	37	40	55	63	65	69	83	97	105	106	110	114	118	122	Mean
BARF	21.4	19.4	18.3	19.3	25.6	3.60	15.1	32.0	1.28	14.9	23.3	19.3	29.8	1.07	22.8	17.8
IDR	27.3	33.5	24.8	21.5	4.02	0.55	0.67	1.66	2.67	0.52	51.2	21.0	22.2	2.70	45.9	17.3
NRF	21.7	25.7	17.9	15.5	11.0	3.20	14.9	23.4	5.28	4.69	42.4	10.6	0.27	15.7	12.4	15.0
NBF*	32.9	26.0	32.6	16.6	1.93	1.45	0.22	0.55	0.59	0.19	26.7	16.2	0.27	0.55	18.8	11.7
GBF	0.08	11.5	9.20	8.94	**0.05**	**0.04**	0.08	25.7	0.23	0.07	39.7	**0.09**	14.6	0.09	18.6	8.61
SPF	**0.07**	0.38	0.08	0.20	0.40	0.14	**0.08**	0.35	**0.17**	0.13	0.10	0.14	0.14	0.18	0.15	0.18
Ours	**0.07**	**0.17**	**0.06**	**0.08**	0.06	0.21	0.10	**0.17**	0.21	**0.06**	**0.06**	0.18	**0.09**	**0.08**	**0.14**	**0.12**

Table 1: The quantitative comparison results in terms of RMSE accuracy (both translation and rotation errors) of camera pose estimation from different comparing approaches on DTU dataset. NBF (NBF*) represents NeuS-BARF (NeuS-BARF*), GBF represents Geo-NeuS-BARF and SPF represents SPARF respectively.

Scan	24	37	40	55	63	65	69	83	97	105	106	110	114	118	122	Mean
BARF	7.62	8.82	8.86	8.34	8.19	6.85	8.58	9.44	7.87	9.93	7.01	7.87	7.96	6.32	7.83	8.10
IDR	8.48	9.21	8.74	10.4	11.4	7.19	3.47	7.32	9.47	4.67	8.12	9.21	7.71	7.21	8.72	8.09
NBF	7.93	7.89	8.51	9.39	9.07	8.53	8.23	8.78	9.77	8.31	9.33	6.35	8.88	7.59	9.68	8.55
NBF*	9.22	9.90	8.31	9.16	9.60	5.94	3.75	7.04	5.64	2.37	8.59	9.00	1.10	3.47	8.67	6.78
GBF	1.54	8.30	8.73	8.98	1.75	2.15	**0.78**	9.88	1.21	1.80	9.49	**0.78**	9.13	0.90	9.65	5.00
SPF	2.10	5.89	2.10	1.57	8.90	**1.81**	1.53	9.39	1.47	3.03	2.18	2.55	0.76	1.20	1.43	3.06
Ours	**1.07**	**2.14**	**1.55**	**1.38**	**1.31**	2.03	0.81	**2.95**	**1.02**	**1.39**	**1.30**	1.62	**0.37**	**0.88**	**1.37**	**1.41**

Table 2: The quantitative comparison results of Chamfer Distance accuracy on DTU dataset. NBF (NBF*) represents NeuS-BARF (NeuS-BARF*), GBF represents Geo-NeuS-BARF and SPF represents SPARF respectively.

Experiments and Analysis

Experimental Settings

Dataset. Firstly, we choose to evaluate our approach on the public DTU dataset (Aanæs et al. 2016) with 15 different object scan. For sparse views, we follow (Long et al. 2022) and (Lin et al. 2021) to randomly select as few as 3 views for each object scan, and then synthetically perturb its camera pose with an additive Gaussian noise $\mathcal{N}(0, 0.15)$. Besides, we also evaluate on 7 challenging scenes from low-res set of the BlendedMVS dataset (Yao et al. 2020), and similarly select 3 views from them.

Baselines. We first choose BARF (Lin et al. 2021), IDR (Yariv et al. 2020) and SPARF (Truong et al. 2023) for comparison. Besides, we also choose NeuS (Wang et al. 2021a) and GeoNeuS (Fu et al. 2022), as with their BARF-based versions, called "NeuS-BARF", "GeoNeuS-BARF" respectively, to enable a fair comparison.

Evaluation on DTU Dataset

Camera Pose Comparison. Table 1 demonstrates the average RMSE accuracy (including both translation and rotation error) between the estimated camera poses and the ground truth camera poses on DTU dataset, using different comparing approaches, including BARF, IDR, NeuS-BARF, GeoNeuS-BARF, SPARF and ours respectively. Note that since NeuS-BARF need extra object mask supervision, we also conduct with it (called as "NeuS-BARF*") for fair comparison. Among all the comparing approaches, the NeRF-like approach BARF, achieves worse RMSE accuracy than the other approaches. This makes sense since other approaches (including our approach) adopt the signed distance field to represent the object's geometry, which is more powerful than the radiance field used in BARF. Although IDR, NeuS-BARF (NeuS-BARF*) and GeoNeuS-BARF achieve various RMSE accuracy in each object scan of DTU dataset respectively, in average they achieve the same level of RMSE accuracy, which means they perform similar camera pose estimation quality. SPARF achieves previous state-of-the-art camera pose estimate accuracy.

In comparison, our approach outperforms SPARF with more better average RMSE accuracy, where we outperform SPARF in 13(15) scenes in terms of translation, and 11(15) scenes (with other 2 scenes are the same) in terms of rotation. Note that SPARF utilize extra depth regularization for

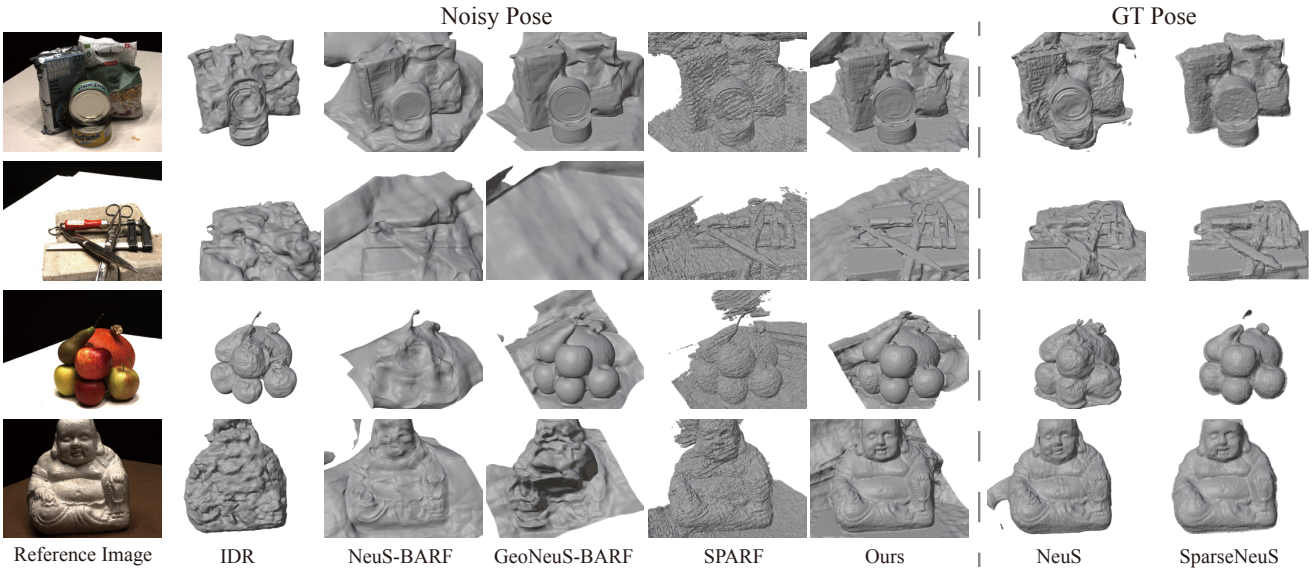

Figure 5: The visual comparing surface reconstruction results using different neural surface reconstruction approaches, with noisy input camera poses (left columns) and ground truth input camera poses (right columns) on DTU dataset.

training while ours didn't. Otherwise, our approach significantly outperforms the other baselines in camera pose estimation. This shows that our SC-NeuS takes effects for camera pose estimation than the other baseline approaches.

Surface Reconstruction Quality. We also compare the surface reconstruction quality between the different comparing approaches. Table 2 demonstrates the quantitative results on Chamfer Distance metric using different approaches evaluated on DTU dataset. Similarly, our approach achieve consistently much better Chamfer Distance accuracy than the other comparing approaches. Fig. 5 illustrates some visual comparison results of the comparing approaches.

Here we didn't include BARF for visual comparison since BARF's camera pose estimation accuracy is significantly worse than the other baselines (as shown in Table 1). Besides, for fair comparison we feed GT pose for NeuS (and SparseNeuS (Long et al. 2022)) in the comparison.

Besides, we can see NeuS-BARF fails to reconstruct fine object surface, which demonstrates that coarse-to-fine position embedding proposed in BARF is not effective to sparse view setting either, even based on NeuS. In contrast, our approach takes effects in joint learning of neural surface representation and camera pose, leading to geometry-consistent surface reconstruction with fine-grained details.

	Translation ↓	Rotation (°) ↓	CD ↓
w/o L_r	22.71	11.77	6.44
w/o L_{ncc}	0.23	0.30	3.44
Full	0.15	0.12	1.41

Table 3: The RMSE (both translation and rotation) and Chamfer Distance (CD) accuracy by different variants of our system.

Please see the fine-grained detail reconstruction by our approach, which is also better than NeuS and SparseNeuS, even with ground truth camera poses as input (Fig. 5). Please refer to supplementary materials for more comparing results.

Evaluation on BlendedMVS Dataset

We also perform evaluation on BlendedMVS dataset to see how our approach behave across different kinds of datasets. Fig. 6 shows some visual comparing surface reconstruction results using different comparing approaches, including NeuS-BARF, IDR, GeoNeuS-BARF, SPARF and our approach. According to the comparison, our approach can achieve much better surface reconstruction quality with fine-grained details than the other approaches. Here we don't include BARF for visual comparison, since BARF fails to converge in most of the comparing cases. Please refer to our supplementary materials [1] for more comparing results.

Ablation and Analysis

View-consistent Re-projection. We first implement a variant version of our full system without using the view-consistent re-projection loss, termed as 'w/o L_r', and perform the surface reconstruction on the DTU dataset. As shown in Table 3, we can see there are large accuracy decrease for both RMSE and CD between 'w/o L_r' and 'Full' systems. This means the view-consistent re-projection loss serves major contribution in our SC-NeuS for the final geometry-consistent surface reconstruction and accurate camera pose estimation. But please note that 'w/o L_r' still outperforms other comparing approaches including BARF, IDR and NeuS-BAFR, by achieving better average RMSE accuracy and CD accuracy in Table 1 and 2.

[1] https://arxiv.org/abs/2307.05892

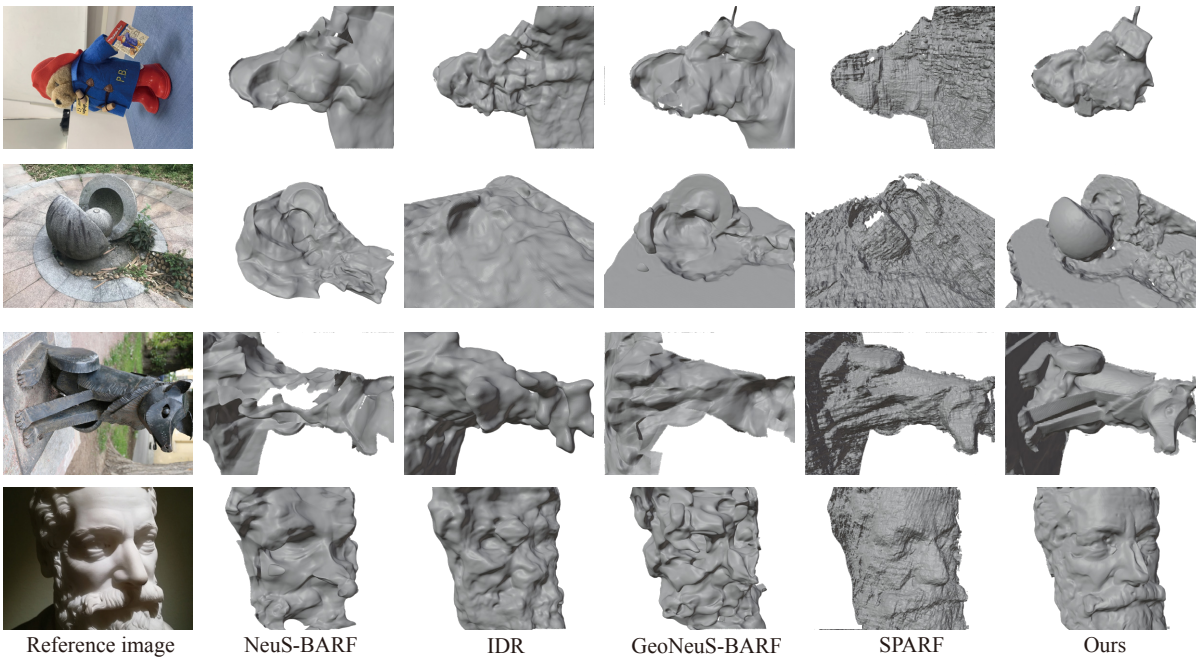

Figure 6: The visual comparing surface reconstruction results using NeuS-BARF, IDR, GeoNeuS-BARF, SPARF and our approach, with sparse view noisy input camera poses on the BlendedMVS dataset.

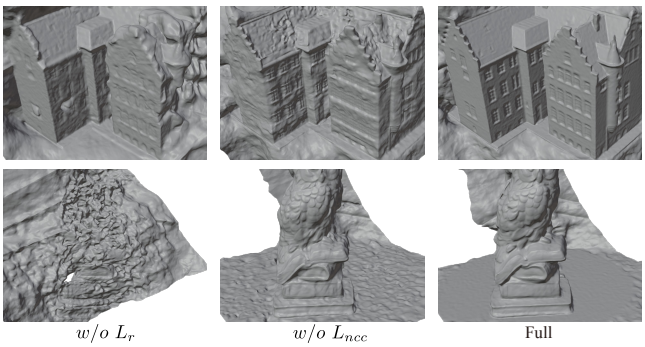

Figure 7: The visual results by different variants of our system.

View-consistent Patch-warping. We also implement a variant system without using the view-consistent patch-warping loss, termed as 'w/o L_{ncc}'. According to Table 3, we can see that 'w/o L_{ncc}' also achieve worse accuracy values than 'Full' in both RMSE for camera pose estimation and CD for surface reconstruction quality, even though the quality decreases are not that much compared with those from 'w/o L_r' to 'Full'.

Fig. 7 shows the visual comparing surface reconstruction results of two example from DTU dataset, using 'w/o L_r', 'w/o L_{ncc}' and 'Full' respectively. We can see there are certain surface quality decrease for our full system ('Full') without using the view-consistent re-projection loss ('w/o L_r'). Even though our approach can achieve fine surface reconstruction without using view-consistent patch-warping loss (see the results of 'w/o L_{ncc}'), we can obvious fine-grained details enhancement by adding the view-consistent loss to our full system (see the results of 'Full'). This means that view-consistent patch-warping loss takes more effective for fine-grained details, while view-consistent re-projection loss works better to boost up the joint learning quality of neural surface and camera pose.

Limitation and Discussion

Our approach's first limitation is that influence from the quality of 2D feature point's matching. Without enough feature point matching in challenging cases like low texture or light changing, our approach couldn't perform well for nice surface reconstruction results. Large camera poses variation between sparse views would also make our approach failed for feasible joint optimization. In the further, we would like to use more robust explicit surface priors for high reliable neural surface reconstruction.

Conclusion

Joint learning for the neural surface representation and camera pose remains to be a challenging problem, especially for sparse scenarios. This paper propose a new joint learning strategy, called SC-NeuS, which explores multi-view constraints directly from the explicit geometry of the neural surface, and achieves consistently better surface reconstruction quality and camera pose estimation accuracy than previous approaches. We hope that our approach can inspire more efforts to the neural surface reconstruction from sparse view images, to enable more feasible real-world applications in this community.

Acknowledgments

This work was supported by Natural Science Foundation of China (Project Number 62202057), the Fundamental Research Funds for the Central Universities (No. 2021NTST07), State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (No. VRLAB2022B03) and CCF- Baidu Open Fund (No. 202304).

References

Aanæs, H.; Jensen, R. R.; Vogiatzis, G.; Tola, E.; and Dahl, A. B. 2016. Large-Scale Data for Multiple-View Stereopsis. *IJCV.*, 120(2): 153–168.

Atzmon, M.; and Lipman, Y. 2020. Sal: Sign agnostic learning of shapes from raw data. In *IEEE CVPR*, 2565–2574.

Azinović, D.; Martin-Brualla, R.; Goldman, D. B.; Nießner, M.; and Thies, J. 2022. Neural rgb-d surface reconstruction. In *IEEE CVPR*, 6290–6301.

Barron, J. T.; Mildenhall, B.; Verbin, D.; Srinivasan, P. P.; and Hedman, P. 2022. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *IEEE CVPR*, 5470–5479.

Boss, M.; Engelhardt, A.; Kar, A.; Li, Y.; Sun, D.; Barron, J. T.; Lensch, H.; and Jampani, V. 2022. SAMURAI: Shape And Material from Unconstrained Real-world Arbitrary Image collections. In *Advances in Neural Information Processing Systems*.

Chen, A.; Xu, Z.; Zhao, F.; Zhang, X.; Xiang, F.; Yu, J.; and Su, H. 2021. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *IEEE CVPR*, 14124–14133.

Chng, S.-F.; Ramasinghe, S.; Sherrah, J.; and Lucey, S. 2022. Gaussian activated neural radiance fields for high fidelity reconstruction and pose estimation. In *ECCV*, 264–280. Springer.

Darmon, F.; Bascle, B.; Devaux, J.-C.; Monasse, P.; and Aubry, M. 2022. Improving neural implicit surfaces geometry with patch warping. In *IEEE CVPR*, 6260–6269.

Deng, K.; Liu, A.; Zhu, J.-Y.; and Ramanan, D. 2022. Depth-supervised nerf: Fewer views and faster training for free. In *IEEE CVPR*, 12882–12891.

Fu, Q.; Xu, Q.; Ong, Y. S.; and Tao, W. 2022. Geo-neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction. *Advances in Neural Information Processing Systems*, 35: 3403–3416.

González, V. M.; Gilbert, A.; Phillipson, G.; Jolly, S.; and Hadfield, S. 2022. SVS: Adversarial refinement for sparse novel view synthesis. *arXiv preprint arXiv:2211.07301*.

Huang, J.; Huang, S.-S.; Song, H.; and Hu, S.-M. 2021. Difusion: Online implicit 3d reconstruction with deep priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8932–8941.

Jensen, R.; Dahl, A.; Vogiatzis, G.; Tola, E.; and Aanæs, H. 2014. Large scale multi-view stereopsis evaluation. In *IEEE CVPR*, 406–413.

Jeong, Y.; Ahn, S.; Choy, C.; Anandkumar, A.; Cho, M.; and Park, J. 2021. Self-calibrating neural radiance fields. In *IEEE CVPR*, 5846–5854.

Jiang, C.; Sud, A.; Makadia, A.; Huang, J.; Nießner, M.; Funkhouser, T.; et al. 2020. Local implicit grid representations for 3d scenes. In *CVPR*, 6001–6010.

Kar, A.; Häne, C.; and Malik, J. 2017. Learning a multi-view stereo machine. *Advances in neural information processing systems*, 30.

Kim, M.; Seo, S.; and Han, B. 2022. Infonerf: Ray entropy minimization for few-shot neural volume rendering. In *IEEE CVPR*, 12912–12921.

Kuang, Z.; Olszewski, K.; Chai, M.; Huang, Z.; Achlioptas, P.; and Tulyakov, S. 2022. NeROIC: neural rendering of objects from online image collections. *ACM Transactions on Graphics (TOG)*, 41(4): 1–12.

Labatut, P.; Pons, J.-P.; and Keriven, R. 2007. Efficient multi-view reconstruction of large-scale scenes using interest points, delaunay triangulation and graph cuts. In *IEEE ICCV*, 1–8. IEEE.

Li, D.; Huang, S.-S.; Shen, T.; and Huang, H. 2023. Dynamic View Synthesis with Spatio-Temporal Feature Warping from Sparse Views. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, 1565–1576. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701085.

Lin, C.-H.; Ma, W.-C.; Torralba, A.; and Lucey, S. 2021. Barf: Bundle-adjusting neural radiance fields. In *IEEE CVPR*, 5741–5751.

Long, X.; Lin, C.; Wang, P.; Komura, T.; and Wang, W. 2022. Sparseneus: Fast generalizable neural surface reconstruction from sparse views. In *ECCV*, 210–227. Springer.

Meng, Q.; Chen, A.; Luo, H.; Wu, M.; Su, H.; Xu, L.; He, X.; and Yu, J. 2021. Gnerf: Gan-based neural radiance field without posed camera. In *IEEE ICCV*, 6351–6361.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Niemeyer, M.; Barron, J. T.; Mildenhall, B.; Sajjadi, M. S.; Geiger, A.; and Radwan, N. 2022. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *IEEE CVPR*, 5480–5490.

Oechsle, M.; Peng, S.; and Geiger, A. 2021. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *IEEE ICCV*, 5589–5599.

Park, J. J.; Florence, P.; Straub, J.; Newcombe, R.; and Lovegrove, S. 2019. Deepsdf: Learning continuous signed distance functions for shape representation. In *IEEE CVPR*, 165–174.

Peng, S.; Niemeyer, M.; Mescheder, L.; Pollefeys, M.; and Geiger, A. 2020. Convolutional occupancy networks. In *ECCV*, 523–540. Springer.

Roessle, B.; Barron, J. T.; Mildenhall, B.; Srinivasan, P. P.; and Nießner, M. 2022. Dense depth priors for neural radiance fields from sparse input views. In *IEEE CVPR*, 12892–12901.

Schonberger, J. L.; and Frahm, J.-M. 2016. Structure-from-motion revisited. In *IEEE CVPR*, 4104–4113.

Schönberger, J. L.; Zheng, E.; Frahm, J.-M.; and Pollefeys, M. 2016. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 501–518. Springer.

Snavely, N.; Seitz, S. M.; and Szeliski, R. 2006. Photo tourism: exploring photo collections in 3D. In *ACM SIGGRAPH*, 835–846.

Sucar, E.; Liu, S.; Ortiz, J.; and Davison, A. J. 2021. iMAP: Implicit mapping and positioning in real-time. In *IEEE CVPR*, 6229–6238.

Trevithick, A.; and Yang, B. 2021. Grf: Learning a general radiance field for 3d representation and rendering. In *IEEE ICCV*, 15182–15192.

Truong, P.; Rakotosaona, M.-J.; Manhardt, F.; and Tombari, F. 2023. Sparf: Neural radiance fields from sparse and noisy poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4190–4200.

Wang, P.; Liu, L.; Liu, Y.; Theobalt, C.; Komura, T.; and Wang, W. 2021a. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. *Advances in Neural Information Processing Systems*, 34: 27171–27183.

Wang, Q.; Wang, Z.; Genova, K.; Srinivasan, P. P.; Zhou, H.; Barron, J. T.; Martin-Brualla, R.; Snavely, N.; and Funkhouser, T. 2021b. Ibrnet: Learning multi-view image-based rendering. In *IEEE CVPR*, 4690–4699.

Wang, Z.; Wu, S.; Xie, W.; Chen, M.; and Prisacariu, V. A. 2021c. NeRF−−: Neural Radiance Fields Without Known Camera Parameters. *arXiv preprint arXiv:2102.07064*.

Wei, Y.; Liu, S.; Rao, Y.; Zhao, W.; Lu, J.; and Zhou, J. 2021. Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In *IEEE CVPR*, 5610–5619.

Xiao, Y.; Xue, N.; Wu, T.; and Xia, G.-S. 2023. Level-S ̃2 fM: Structure From Motion on Neural Level Set of Implicit Surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17205–17214.

Xu, Q.; and Tao, W. 2019. Multi-scale geometric consistency guided multi-view stereo. In *IEEE CVPR*, 5483–5492.

Xu, Q.; and Tao, W. 2020. Pvsnet: Pixelwise visibility-aware multi-view stereo network. *arXiv preprint arXiv:2007.07714*.

Yao, Y.; Luo, Z.; Li, S.; Fang, T.; and Quan, L. 2018. Mvsnet: Depth inference for unstructured multi-view stereo. In *ECCV*, 767–783.

Yao, Y.; Luo, Z.; Li, S.; Zhang, J.; Ren, Y.; Zhou, L.; Fang, T.; and Quan, L. 2020. BlendedMVS: A Large-Scale Dataset for Generalized Multi-View Stereo Networks. In *IEEE CVPR*, 1787–1796.

Yariv, L.; Gu, J.; Kasten, Y.; and Lipman, Y. 2021. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34: 4805–4815.

Yariv, L.; Kasten, Y.; Moran, D.; Galun, M.; Atzmon, M.; Ronen, B.; and Lipman, Y. 2020. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33: 2492–2502.

Yu, A.; Ye, V.; Tancik, M.; and Kanazawa, A. 2021. pixelnerf: Neural radiance fields from one or few images. In *IEEE CVPR*, 4578–4587.

Zhang, J.; Yang, G.; Tulsiani, S.; and Ramanan, D. 2021. NeRS: neural reflectance surfaces for sparse-view 3D reconstruction in the wild. *Advances in Neural Information Processing Systems*, 34: 29835–29847.

Zhang, J. Y.; Ramanan, D.; and Tulsiani, S. 2022. Relpose: Predicting probabilistic relative rotation for single objects in the wild. In *ECCV*, 592–611. Springer.

Zhu, Z.; Peng, S.; Larsson, V.; Xu, W.; Bao, H.; Cui, Z.; Oswald, M. R.; and Pollefeys, M. 2022. Nice-slam: Neural implicit scalable encoding for slam. In *IEEE CVPR*, 12786–12796.

Zou, Z.-X.; Huang, S.-S.; Mu, T.-J.; and Wang, Y.-P. 2022. ObjectFusion: Accurate object-level SLAM with neural object priors. *Graphical Models*, 123: 101165.