NeuSurf: On-Surface Priors for Neural Surface Reconstruction from Sparse Input Views

Han Huang[1,2], **Yulun Wu**[1,2], **Junsheng Zhou**[1,2], **Ge Gao**[1,2*], **Ming Gu**[1,2], **Yu-Shen Liu**[2]

[1]Beijing National Research Center for Information Science and Technology (BNRist), Tsinghua University, Beijing, China
[2]School of Software, Tsinghua University, Beijing, China
{h-huang20, wu-yl22, zhoujs21}@mails.tsinghua.edu.cn, {gaoge, guming, liuyushen}@tsinghua.edu.cn

Abstract

Recently, neural implicit functions have demonstrated remarkable results in the field of multi-view reconstruction. However, most existing methods are tailored for dense views and exhibit unsatisfactory performance when dealing with sparse views. Several latest methods have been proposed for generalizing implicit reconstruction to address the sparse view reconstruction task, but they still suffer from high training costs and are merely valid under carefully selected perspectives. In this paper, we propose a novel sparse view reconstruction framework that leverages on-surface priors to achieve highly faithful surface reconstruction. Specifically, we design several constraints on global geometry alignment and local geometry refinement for jointly optimizing coarse shapes and fine details. To achieve this, we train a neural network to learn a global implicit field from the on-surface points obtained from SfM and then leverage it as a coarse geometric constraint. To exploit local geometric consistency, we project on-surface points onto seen and unseen views, treating the consistent loss of projected features as a fine geometric constraint. The experimental results with DTU and BlendedMVS datasets in two prevalent sparse settings demonstrate significant improvements over the state-of-the-art methods.

Introduction

Surface reconstruction from multi-view images is a critical task in the fields of computer vision and computer graphics. Traditional methods, like Multi-View Stereo (Campbell et al. 2008; Schonberger and Frahm 2016; Yao et al. 2018), leverage geometric consistency between images to compute the depth map. Subsequently, they obtain the reconstructed point cloud through depth map fusion. Nonetheless, the conversion of this intermediate representation might introduce cumulative geometric errors. In scenarios with sparse views, the MVS method faces challenges in reconstructing a smooth and detailed surface due to the scarcity of matching points and variations in viewing angles.

In recent years, neural rendering-based surface reconstruction methods (Yariv et al. 2021; Wang et al. 2021; Yu et al. 2022) have been widely used to improve the reconstruction results by producing smoother and more complete geometries. These methods simultaneously optimize both implicit geometry and neural radiance fields by minimizing the discrepancy between the rendered and the ground truth images. However, the well-optimized photometric loss may distort the geometry due to shape-radiance ambiguity (Zhu et al. 2023). Especially under the situation of sparse view input, learning of the geometry field may further collapse.

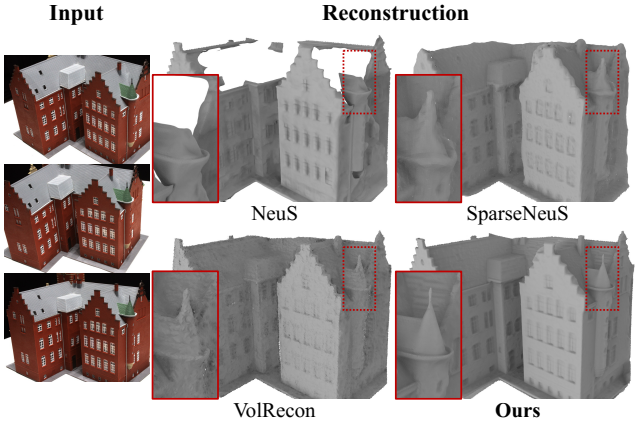


Figure 1: Surface reconstruction results from sparse-view images (large-overlap setting). The state-of-the-art methods SparseNeuS (Long et al. 2022) and VolRecon (Ren et al. 2023) produce noisy and broken surfaces, while the results of ours (NeuSurf) are detailed and complete.

As a remedy, some recent methods (Long et al. 2022; Ren et al. 2023) partly solved the sparse view reconstruction by introducing additional generalizable priors. They first learn geometric priors from large-scale data, and then fine-tune the implicit fields to achieve surface reconstruction in new scenarios. However, the learned priors are only effective in a fixed large-overlap sparse setting. Once the sparse view is inconsistent with the pose distribution with the fixed setting, the priors will be invalid and fail to bring robust guidance to surface reconstruction. As a result, the performance of the generalizable methods is dramatically limited in cases with different sparse settings. Due to the long training time and complex data pre-processing, it is unrealistic to train a prior for each sparse setting.

In this paper, we propose a novel sparse view reconstruc-

*Corresponding author.

tion framework to achieve highly faithful surface reconstructions using on-surface point priors. The proposed priors are achieved directly from the raw input sparse views without requiring any extra training or data, which effectively improve the reconstruction results and are robust to different sparse settings. The effectiveness of our method is not affected by different sparse settings. Specifically, we obtain initial on-surface points by SfM method, which can be considered as "free data" in the training process. Instead of using these points directly for depth supervision, we design two constraints in terms of loss functions to make full use of on-surface points. One is conducted with the guidance of the global geometry field, where we train a neural network to learn the geometric field of on-surface points and then use that field as a rough geometric constraint. The other is local geometric refinement loss, which is achieved by projecting surface points onto visible and invisible views and optimizing the consistency of projection features to reconstruct fine local geometry. Our contributions are listed below.

- We propose a novel framework for surface reconstruction from sparse view images. our framework takes full advantage of on-surface point clouds, which is easy to access, as an additional effective supervision to guide the geometry learning.
- We use the global geometric fields obtained from the surface points to help learn rough and continuous geometry. In addition, we optimize the local feature consistency of on-surface points to help learn fine geometry.
- We achieve state-of-the-art reconstruction results under different prevalent sparse settings on the widely-used DTU and BlendedMVS datasets.

Related Work

Multi-View Stereo (MVS)

Traditionally, MVS methods use point clouds (Furukawa and Ponce 2010; Lhuillier and Quan 2005), depth maps (Galliani, Lasinger, and Schindler 2015; Schönberger et al. 2016; Xu and Tao 2019) and voxel grids (Kostrikov, Horbert, and Leibe 2014; Choe et al. 2021; Ji et al. 2017) as 3D representations of scenes to reconstruct the surface geometry. Due to the need for parallel computing, depth maps based methods are now widely used. Depth maps based methods predict the depth of each image and then fuse them to get the surface point cloud of the object. After the point cloud is obtained, Screened Poisson surface reconstruction method (Kazhdan and Hoppe 2013) can be used to further obtain the surface mesh.

Neural Implicit Representations

Recently, advanced methods employing neural implicit functions to represent 3D scenes have emerged, and these can be applied to shape representation (Zhou et al. 2022b; Mescheder et al. 2019; Zhou et al. 2023), novel view synthesis (Mildenhall et al. 2020; Liu et al. 2020; Zhang et al. 2023) and multi-view 3D reconstruction (Oechsle, Peng, and Geiger 2021; Wang et al. 2021; Yariv et al. 2021; Yu et al. 2022; Ma et al. 2023). Given raw point clouds, Neural-pull (Ma et al. 2021) and CAP-UDF (Zhou et al. 2022a) design neural networks to learn geometric fields that represent 3D shapes. They provide a way to transform the raw point cloud representation of an object surface into a geometric field representation without the ground truth values of the geometric field. Neural Radiance Fields (NeRF) (Mildenhall et al. 2020), as a popular novel view synthesis method in recent years, encodes color fields and volume density fields with implicit representations.

Neural Surface Reconstruction

Inspired by NeRF, NeuS (Wang et al. 2021) and VolSDF (Yariv et al. 2021) were proposed to encode signed distance function (SDF) and color fields of the scene. By minimizing the discrepancy between the rendered image and the ground truth image, they can obtain a smooth and complete SDF geometric field. To make the geometry of the SDF field more precise, MonoSDF (Yu et al. 2022) and Geo-NeuS (Fu et al. 2022) add geometric loss in addition to photometric loss, which reduces the possible bias in the volume rendering process. The above-mentioned methods are all based on dense view input. However, in real world, there are often only fewer views that can be used for reconstruction.

To achieve sparse view reconstruction, SparseNeuS (Long et al. 2022) and VolRecon (Ren et al. 2023) learn generalizable geometric priors from large-scale data, and then finetune on new scenes. They train on 75 scenes of the DTU (Jensen et al. 2014) for several days and then test on the remaining 15 scenes. Although they have obtained some geometric priors of the data sets through large-scale training, they still only generalize in the case of a fixed sparse setting.

In observation, when only sparse views are given, the complexity of the neural surface learning increases, and it is more likely to achieve a collapsed geometry (incomplete, noisy), as shown in Figure 1. Current neural surface learning methods with large-scale training priors are often time-intensive and only useful within a specific sparse setting.

Unlike previous works, instead of using costly large-training priors, our method attempts to exploit the priors of surface points to optimize the neural surface representation both globally and locally.

Method

Given sparse-view images $I = \{I_i | i \in 1, .., \mathcal{M}\}$ with camera poses $T = \{T_i, | i \in 1, ..., \mathcal{M}\}$, our goal is to reconstruct the high-quality geometry S of the scene represented by I. In this paper, we propose NeuSurf, a neural surface reconstruction method with sparse input views, as illustrated in Figure 2.

Our motivation for proposing NeuSurf is to reduce the complexity of neural surface learning with non-training priors and produce more complete and detailed reconstruction. Specifically, points obtained by Structure from Motion (Schonberger and Frahm 2016) are regarded as a "free" data source as it is easy to acquire with no extra input. We denote the surface points obtained by SfM as P, which we do not use as a depth loss function for the neural rendering directly. Instead, we learn a global geometric field f_θ from these on-surface points and use it to align the rough geometry. To get

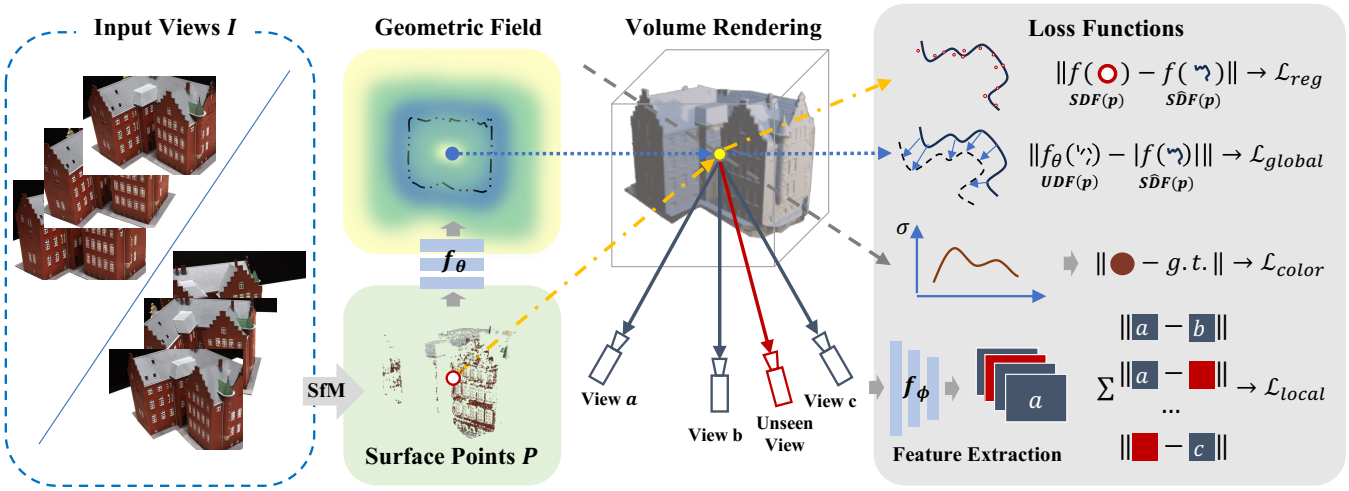


Figure 2: Structure of NeuSurf. For a set of 3 source views (in large-overlap or little-overlap), we first obtain the surface points by the SfM method. Within the on-surface points, we train a UDF network as the geometric field and leverage it as global geometry alignment. Then we utilize the feature consistency between seen and unseen views to optimize the local geometry. In addition to the RGB rendering loss, explicit on-surface points regularization can be improved as an additional loss.

fine details of the surface, local feature consistency for on-surface points is optimized as another constraint.

Learning Neural Implicit Surface by Volume Rendering

We represent the geometry and appearance with SDF fields and color fields, which are learned by the neural rendering pipeline. We adopt NeuS (Wang et al. 2021) as our baseline, which defines the geometry as the zero-level set of signed distance function (SDF) $S = \{\mathbf{x} \in \mathbb{R}^3 | f(\mathbf{x}) = 0\}$, and develop a novel volume rendering method to learn the geometry and appearance of the scene. The SDF and color are parameterized with two MLPs as provided by NeuS.

Given a pixel from one image, the ray could be denoted as $\{\mathbf{r}(t) = \mathbf{o} + t\mathbf{d} | t > 0\}$, where $\mathbf{o}$ is the camera center and $\mathbf{d}$ is the direction of the ray. The rendered color is accumulated by volume rendering with N discrete points:

$$C(\mathbf{r}) = \sum_{i=1}^{N} T_i \alpha_i c_i, \tag{1}$$

where T_i is the accumulated transmittance, α_i is opacity values, as denoted by

$$T_i = \prod_{j=1}^{i-1} (1 - \alpha_i), \tag{2}$$

$$\alpha_i = \max\left(\frac{\Phi_s(f(\mathbf{r}(t_i))) - \Phi_s(f(\mathbf{r}(t_{i+1})))}{\Phi_s(f(\mathbf{r}(t_i)))}, 0 \right). \tag{3}$$

Φ_s follows NeuS, expressed as $\Phi_s(x) = (1 + e^{-sx})^{-1}$ with s being a trainable, diminishing parameter.

On-Surface Global Geometry Alignment

On-surface points are discrete and thus fail to determine some surface locations. We attempt to learn a continuous geometric field from the discrete point cloud as a prior to provide coarse information for surface learning. With the guidance of the prior field, we largely reduce the difficulties in optimization with neural volume rendering, thus enabling robust learning process. We also justify that the prior is the key factor that prevents collapse in the difficult sparse-view setting, we are able to obtain a rough but relatively complete geometric surface.

An intuitive solution is first to train an SDF Network for on-surface points and then use it as a part of NeuS directly. This is like pre-training the SDF function by on-surface points and then fine-tuning it with 2D images. However, fitting an SDF field to a sparse point cloud is challenging, due to the geometry complexity and the closed surface assumption. Even though our reconstruction targets are all closed surfaces, in sparse view, the on-surface points computed by the SfM method are only part of the surface of the objects. It means that enforcing the closed surface assumption leads to the overfitting of the geometric field.

To resolve this issue, the unsigned distance function (UDF) is considered to fit the global geometric field of the on-surface points. It is flexible and can cope with open surfaces. We train a UDF network f_θ to fit the surface points to obtain a complete and continuous geometric field. With the geometric field, we treat it as a coarse global prior that can stabilize the geometry optimizing with neural renderings.

Since the ground truth UDF values are not provided, we follow the CAP-UDF (Zhou et al. 2022a) to train a network f_θ within a moving operation. We randomly sample a set of query locations $Q = \{\mathbf{q}_i, i \in [1, M]\}$ around given on-surface points P. Then we move the point $\mathbf{q}_i$ against the direction of the gradient at $\mathbf{q}_i$ with a stride of UDF value $f_\theta(\mathbf{q}_i)$. Since the gradient points to the steepest uphill direction, the moving operation aims to find a path to pull point

$\mathbf{p}_i$ onto the surface S. The operation can be formulated as:

$$\mathbf{z}_i = \mathbf{q}_i - f_\theta(\mathbf{q}_i) \times \nabla f_\theta(\mathbf{q}_i)/||\nabla f_\theta(\mathbf{q}_i)||_2. \quad (4)$$

Here, $\mathbf{z}_i$ is the location after the moving operation. The moving operation is differentiable in both the unsigned distance value and the gradient, which allows us to optimize them simultaneously during training.

For a well-learned network f_θ, the moved points should be on the surface, which can be used as our learning objective. Hence, the Chamfer Distance between the moved points and the on-surface points can be used as $\mathcal{L}_{udf}$ to optimize our UDF network f_θ:

$$\mathcal{L}_{udf} = \mathrm{CD}(Z, Q). \quad (5)$$

With a trained UDF network f_θ, we attempt to incorporate the knowledge of the continuous geometry field into neural rendering. For the effect of network f_θ, query points closer to the surface are more accurate. Therefore, we design a cut-off threshold ϵ for network f_θ to supervise the geometric field near the surface. The global geometry alignment is given by

$$\mathcal{L}_{global} = \frac{1}{|N|} \sum_{\mathbf{n}_i \in N} |f(\mathbf{n}_i)| \left(1 - \frac{\max(f_\theta(\mathbf{n}_i) - \epsilon, 0)}{f_\theta(\mathbf{n}_i) - \epsilon}\right), \quad (6)$$

where N and ϵ are the discrete points in ray rendering and on-surface alignment threshold value, respectively.

On-Surface Local Geometry Refinement

Geometric field alignment can constrain the shape and ensure the integrity of the reconstructed object. However, the learned geometric field is coarse and can not hone the details of the reconstruction. The reason is that the sparse input provides less surface feature information, this may lead to severe noise. Therefore, a local-level optimization is needed. Our insight comes from the traditional MVS methods (Yao et al. 2018; Ji et al. 2017) where the correctness of a surface point estimation is guaranteed by the consistency of its corresponding feature between different views. We justify that on-surface points obtained by SfM methods also follow this assumption. Inspired by that, we render images with given poses and a novel pose. By supervising the projection features of the on-surface points, geometry and color fields are optimized simultaneously, which can be expressed by

$$\mathcal{L}_{local} = \frac{1}{|P||I|} \sum_{\mathbf{p}_i \in P} \sum_{\mathcal{H}_j \in \mathcal{H}} \left\| f_\phi(\mathcal{H}_j(\mathbf{p}_i)) - f_\phi(\mathcal{H}_0(\mathbf{p}_i)) \right\|, \quad (7)$$

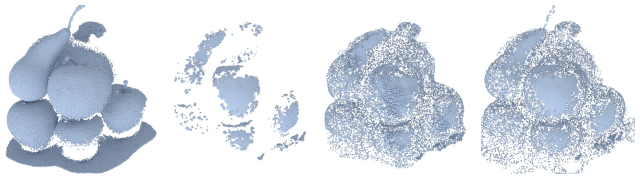
where P is on-surface point cloud obtained by SfM method, f_ϕ is a geometry feature extraction network, I and $\mathcal{H}$ are input views and the transformation matrices respectively.

However, initial points obtained by SfM method are limited in quantity and unevenly distributed. The points are barely concentrated on positions of relatively poor visibility, which leads to biased optimization spatially. Hence, during the ray rendering process, we obtain some points by calculating the ray-surface intersection points P'. We name them pseudo on-surface points since they are acquired from the implicit surface to be optimized. The specific point $\mathbf{p}'$ passed through by a ray r during neural rendering is denoted as $\mathbf{p}' = \{\mathbf{r}(t^*) | f(\mathbf{r}(t^*)) = 0\}$. We first find t_i by solving $f(\mathbf{r}(t_i))f(\mathbf{r}(t_{i+1})) < 0$. And t^* can be calculated as:

$$t^* = \frac{f(\mathbf{r}(t_i))t_{i+1} - f(\mathbf{r}(t_{i+1}))t_i}{f(\mathbf{r}(t_i)) - f(\mathbf{r}(t_{i+1}))}. \quad (8)$$

Since the computation of the pseudo on-surface points is differentiable, it is reasonable to optimize the surface by considering them within the feature projection alignment loss. The implicit surface together with 3D coordinates of the pseudo on-surface points themselves are optimized along the training process, as we can see in Figure 3. Therefore, we add differentiable pseudo surface points P' to surface points P during the training process to keep the projected features consistent between seen and unseen views:

$$P \leftarrow P + P'. \quad (9)$$



(a) (b) (c) (d)

Figure 3: (a) Ground truth on-surface points; (b) On-surface points generated with SfM; (c) Pseudo and SfM on-surface points at early stages of training; (d) Pseudo and SfM on-surface points at the end of training process. The pseudo on-surface points are optimized to potential real surface.

Loss Functions

The overall loss functions are:

$$\mathcal{L} = \mathcal{L}_{color} + \lambda_1 \mathcal{L}_{global} + \lambda_2 \mathcal{L}_{local} + \lambda_3 \mathcal{L}_{eik} + \lambda_4 \mathcal{L}_{reg}, \quad (10)$$

where $\mathcal{L}_{global}$ and $\mathcal{L}_{local}$ are the on-surface global geometry alignment loss and local refinement loss defined above.

$\mathcal{L}_{color}$ is the difference between the rendered and ground-truth pixel colors:

$$\mathcal{L}_{color} = \frac{1}{\mathcal{R}} \sum_{\mathbf{r} \in \mathcal{R}} ||C(\mathbf{r}) - \hat{C}(\mathbf{r})||. \quad (11)$$

As with NeuS (Wang et al. 2021), an Eikonal term (Gropp et al. 2020) on the random sample points $\mathcal{Y}$ to regularize SDF of $f(x)$ is introduced:

$$\mathcal{L}_{eik} = \frac{1}{|\mathcal{Y}|} \sum_{\mathbf{x} \in \mathcal{Y}} (||\nabla f(\mathbf{x})|| - 1)^2. \quad (12)$$

Similar to Geo-NeuS (Fu et al. 2022), AutoRecon (Wang et al. 2023). We adopt $\mathcal{L}_{reg}$ as supervision with a zero-level set. However, given sparse views, the points we obtain by SfM are sparse and precious. We do not supervise rendered depth with real depth for each view. Alternatively, We supervise the SDF values of all points in 3D space in each iteration of training:

$$\mathcal{L}_{reg} = \frac{1}{P} \sum_{\mathbf{p}_i \in P} ||f(\mathbf{p}_i)||. \quad (13)$$

Scan ID	24	37	40	55	63	65	69	83	97	105	106	110	114	118	122	Mean
Little-overlap (PixelNeRF Setting)																
COLMAP	2.88	3.47	**1.74**	2.16	2.63	3.27	2.78	3.63	3.24	3.49	2.46	1.24	1.59	2.72	1.87	2.61
SparseNeuS$_{ft}$	4.81	5.56	5.81	2.68	3.30	3.88	2.39	2.91	3.08	2.33	2.64	3.12	1.74	3.55	2.31	3.34
VolRecon	3.05	4.45	3.36	3.09	2.78	3.68	3.01	2.87	3.07	2.55	3.07	2.77	1.59	3.44	2.51	3.02
NeuS	4.11	5.40	5.10	3.47	2.68	2.01	4.52	8.59	5.09	9.42	2.20	4.84	0.49	2.04	4.20	4.28
VolSDF	4.07	4.87	3.75	2.61	5.37	4.97	6.88	3.33	5.57	2.34	3.15	5.07	1.20	5.28	5.41	4.26
MonoSDF	3.47	3.61	2.10	1.05	2.37	**1.38**	1.41	1.85	1.74	1.10	1.46	2.28	1.25	1.44	1.45	1.86
Ours	**1.35**	**3.25**	2.50	**0.80**	**1.21**	2.35	**0.77**	**1.19**	**1.20**	**1.05**	**1.05**	**1.21**	**0.41**	**0.80**	**1.08**	**1.35**
Large-overlap (SparseNeuS Setting)																
COLMAP	0.90	2.89	1.63	1.08	2.18	1.94	1.61	1.30	2.34	1.28	1.10	1.42	0.76	1.17	1.14	1.52
SparseNeuS$_{ft}$	2.17	3.29	2.74	1.67	2.69	2.42	1.58	1.86	1.94	1.35	1.50	1.45	0.98	1.86	1.87	1.96
VolRecon	1.20	2.59	1.56	1.08	1.43	1.92	1.11	1.48	1.42	1.05	1.19	1.38	0.74	1.23	1.27	1.38
NeuS	4.57	4.49	3.97	4.32	4.63	1.95	4.68	3.83	4.15	2.50	1.52	6.47	1.26	5.57	6.11	4.00
VolSDF	4.03	4.21	6.12	0.91	8.24	1.73	2.74	1.82	5.14	3.09	2.08	4.81	0.60	3.51	2.18	3.41
MonoSDF	2.85	3.91	2.26	1.22	3.37	1.95	1.95	5.53	5.77	1.10	5.99	2.28	0.65	2.65	2.44	2.93
Ours	**0.78**	**2.35**	**1.55**	**0.75**	**1.04**	**1.68**	**0.60**	**1.14**	**0.98**	**0.70**	**0.74**	**0.49**	**0.39**	**0.75**	**0.86**	**0.99**

Table 1: The quantitative comparison results of Chamfer Distances (CD↓) on DTU dataset.

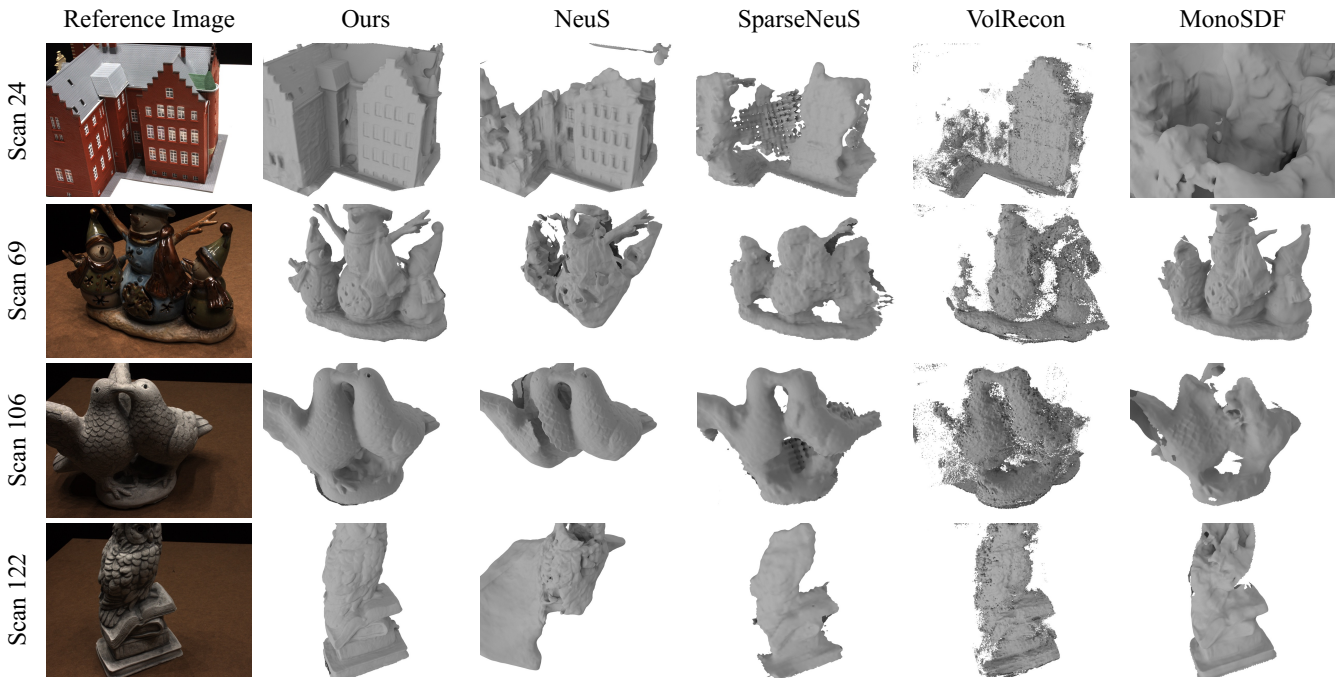


Figure 4: Visual comparisons on the little-overlap sparse setting of DTU dataset. (*NeuS cannot generate valid mesh for scan 122 with the generic 3 views. We take an additional view for training on this scan with NeuS for visual comparison.)

Experiments and Analysis

We conduct abundant experiments on several generic public MVS datasets (Jensen et al. 2014; Yao et al. 2020) from sparse views. We compare our results with some recently presented neural implicit surface reconstruction methods, including the previous state-of-the-art sparse views specified methods. Then we give an ablation study of our approach.

Experimental Settings

Datasets. Previous neural sparse view reconstruction approaches normally select 3 proper input views from each scan of the DTU dataset (Jensen et al. 2014), which contains from 49 to 64 images at a resolution of 1200×1600 for each object scan with known camera intrinsics and poses, to evaluate the performance of the models. We notice that

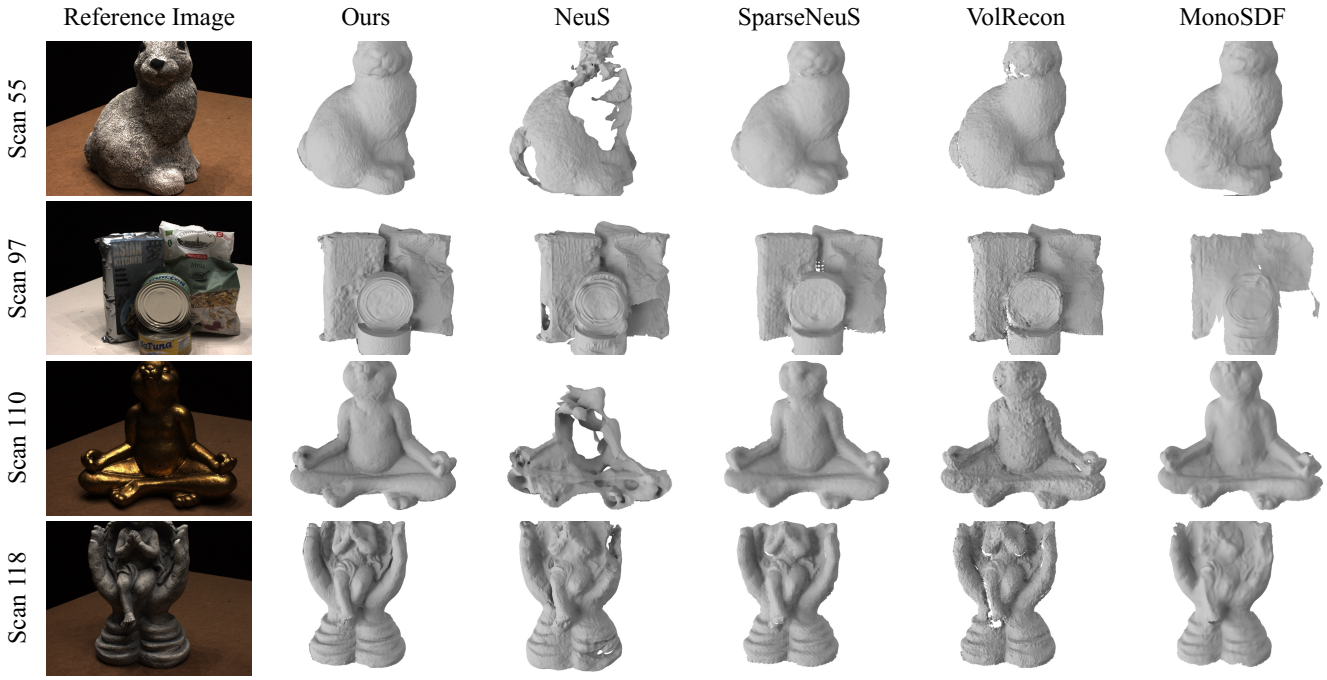


Figure 5: Visual comparisons of surface reconstruction results on the large-overlap sparse setting of DTU dataset.

different approaches differ in the choice of concrete input views. SparseNeus (Long et al. 2022) and VolRecon (Ren et al. 2023) take views 23, 24 and 33 of each scan as one of the test sets for three-view reconstruction. We name it *large-overlap* setting because the distribution of the selected views is concentrated and the visibility overlap between the pics is relatively large. While MonoSDF (Yu et al. 2022) follows PixelNeRF (Yu et al. 2022), taking views 22, 25 and 28 of each scan as sparse-view setting, which we name *little-overlap* setting because of the scattered view distribution. To evaluate our approach and the baselines comprehensively and fairly, we conduct experiments on both two settings on 15 scans commonly used for evaluation.

Besides, we also employ the BlendedMVS dataset (Yao et al. 2020) to estimate the versatility of our approach. We randomly select 3 views from each scene as input and conduct the evaluation on 8 challenging scenes at a resolution of 768×576.

Baselines. We compare our approach with various types of surface reconstruction methods on adopted datasets. **a.** COLMAP (Schonberger and Frahm 2016): A widely used classical SfM framework, which is also the pro-precessing approach we employ in our pipeline. **b.** Generalizable neural implicit surface reconstruction methods, including SparseNeuS$_{ft}$ (Long et al. 2022) and VolRecon (Ren et al. 2023). ft indicates that we do fine-tuning on every single scene before we test the model. **c.** Per-scene optimization methods, including NeuS (Wang et al. 2021), VolSDF (Yariv et al. 2021) and MonoSDF (Yu et al. 2022). We adjust the experiment settings for specific baselines to maximize their performance.

Implementation details. We use naive COLMAP (Schonberger and Frahm 2016) for feature matching to obtain the coarsely estimated point clouds of the test scenes with ground truth poses as inputs. We implement SDF representation model and neural radiance field based on NeuS (Wang et al. 2021) baseline and adopt similar network architecture as CAP-UDF (Zhou et al. 2022a) to learn UDF network f_θ. To achieve better local geometry refinement, we use Vis-MVSNet (Zhang et al. 2020) as the feature extraction network f_ϕ.

For a training procedure of a single scene, we sample 512 rays per batch and train the model for 300k iterations on an NVIDIA RTX3090 GPU.

Comparisons

Sparse View Reconstruction on DTU. We conduct comparisons on both two DTU sparse settings without mask supervision. We measure the Chamfer Distances on DTU dataset in the same way as (Ren et al. 2023) to quantitatively evaluate the reconstruction quality, which is demonstrated in Table 1. Our approach achieves better performance on most of the scenes in little-overlap setting and all of the scenes in large-overlap setting, outperforming the compared baselines including previous state-of-the-art methods.

We present visualizations for both types of DTU sparse-settings. Figure 4 shows the reconstruction results on little-overlap setting, while Figure 5 shows the reconstruction results of the approaches on large-overlap setting. It is challenging for most of the compared methods to obtain complete geometry when the distribution of input views is relatively concrete, while our approach not only captures enough global information to rebuild the correct coarse

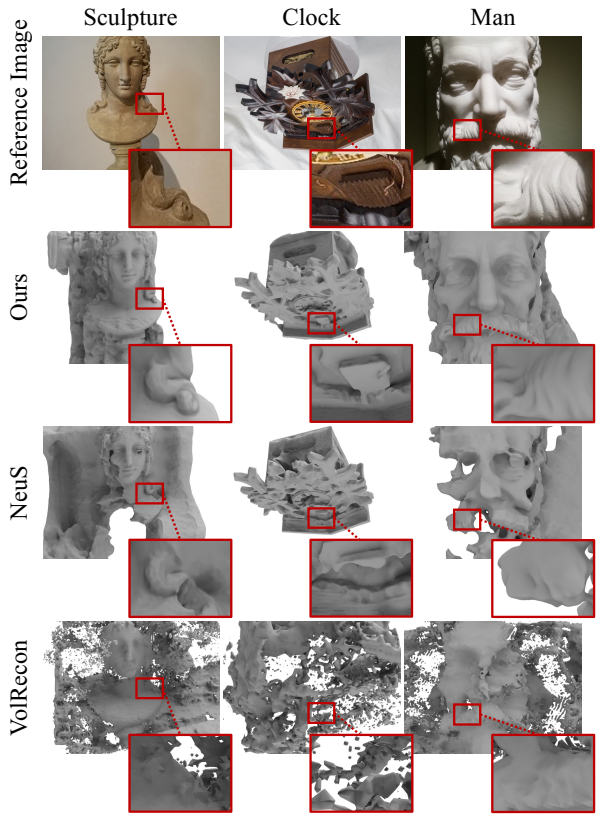


Figure 6: Visual comparison of surface reconstruction results on BlendedMVS dataset.

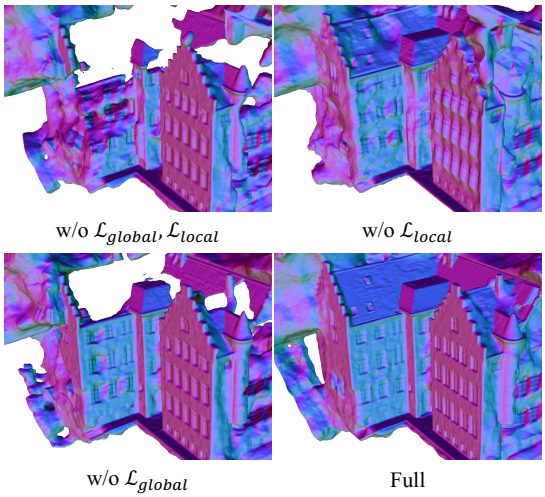


Figure 7: Comparison for reconstructed normal maps of ablation results on DTU scan 24.

$\mathcal{L}_{global}$	$\mathcal{L}_{local}$	Mean CD↓
×	×	2.46
×	√	1.96
√	×	1.67
√	√	**1.35**

Table 2: Reconstruction results comparison of mean Chamfer Distance on little-overlap sparse input subset of DTU dataset by variants of our approach.

shape but also better restore the facial details of the objects.

Sparse View Reconstruction on BlendedMVS. To evaluate the generalization ability of our approach on different datasets, we perform an evaluation on BlendedMVS dataset. We conduct reconstruction tests on 8 representative scenes, from each of which 3 views are randomly selected together with the corresponding camera poses. Some of the reconstruction results are visualized in Figure 6. It shows that our approach could obtain both better global shapes and finer geometric details.

Ablation and Analysis

On-surface global geometric loss $\mathcal{L}_{global}$ and local geometric loss $\mathcal{L}_{local}$ serve as two main components of our reconstruction approach. To better evaluate the effectiveness of these supervisions, we conduct an ablation study. We test our models on the little-overlap DTU sparse setting because this would better reflect the performance of the methods. We separately evaluate the model without $\mathcal{L}_{global}$, the model without $\mathcal{L}_{local}$, and the model without both two losses on all 15 scenes. The mean Chamfer Distances are demonstrated in Table 2.

Although even the model without $\mathcal{L}_{global}$ and $\mathcal{L}_{local}$ still outperforms NeuS baseline, global geometric prior and local feature projection boost the performance to a huge extent. To point out the concrete contributions of these components more intuitively, we give a visualization of the reconstructed normal maps of these ablation models on a single scene in Figure 7. As we can see from the comparison, the rarity of input views leads to the dislocation of some local structures. The participation of $\mathcal{L}_{local}$ alleviates the error. The sparsity of view distribution also introduces a new problem: some spatially continuous parts are incomplete out of the hardness to distinguish foreground and background. Global UDF prior significantly improves the integrity, even when the point clouds obtained by COLMAP are fragmented.

Conclusion

In this paper, we proposed NeuSurf, a novel sparse view surface reconstruction method with on-surface priors. To obtain a rough and complete geometric surface, we train a UDF network to learn the on-surface geometry field and leverage it as global geometry alignment. Then we optimize the feature consistency as local geometry refinement loss to reconstruct detailed surfaces. Our method does not require large-scale training and is robust in various sparse settings. Our method achieves state-of-the-art performance on the DTU dataset in both large-overlap and little-overlap settings. Additionally, we conduct qualitative experiments on the BlendedMVS dataset in different sparse settings and find significant improvement over previous methods.

Acknowledgments

This work was supported by the Program of Science and Technology Plan of Beijing (Z231100001723014).

References

Campbell, N. D.; Vogiatzis, G.; Hernández, C.; and Cipolla, R. 2008. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part I 10*, 766–779. Springer.

Choe, J.; Im, S.; Rameau, F.; Kang, M.; and Kweon, I.-S. 2021. VolumeFusion: Deep Depth Fusion for 3D Scene Reconstruction. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 16066–16075.

Fu, Q.; Xu, Q.; Ong, Y. S.; and Tao, W. 2022. Geo-Neus: Geometry-Consistent Neural Implicit Surfaces Learning for Multi-view Reconstruction. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 3403–3416. Curran Associates, Inc.

Furukawa, Y.; and Ponce, J. 2010. Accurate, Dense, and Robust Multiview Stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8): 1362–1376.

Galliani, S.; Lasinger, K.; and Schindler, K. 2015. Massively Parallel Multiview Stereopsis by Surface Normal Diffusion. In *2015 IEEE International Conference on Computer Vision (ICCV)*, 873–881.

Gropp, A.; Yariv, L.; Haim, N.; Atzmon, M.; and Lipman, Y. 2020. Implicit Geometric Regularization for Learning Shapes. In *Proceedings of Machine Learning and Systems 2020*, 3569–3579.

Jensen, R.; Dahl, A.; Vogiatzis, G.; Tola, E.; and Aanæs, H. 2014. Large Scale Multi-view Stereopsis Evaluation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 406–413.

Ji, M.; Gall, J.; Zheng, H.; Liu, Y.; and Fang, L. 2017. SurfaceNet: An End-To-End 3D Neural Network for Multiview Stereopsis. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2307–2315.

Kazhdan, M.; and Hoppe, H. 2013. Screened Poisson Surface Reconstruction. *ACM Trans. Graph.*, 32(3).

Kostrikov, I.; Horbert, E.; and Leibe, B. 2014. Probabilistic Labeling Cost for High-Accuracy Multi-view Reconstruction. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 1534–1541.

Lhuillier, M.; and Quan, L. 2005. A quasi-dense approach to surface reconstruction from uncalibrated images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3): 418–433.

Liu, L.; Gu, J.; Zaw Lin, K.; Chua, T.-S.; and Theobalt, C. 2020. Neural Sparse Voxel Fields. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 15651–15663. Curran Associates, Inc.

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

Ma, B.; Han, Z.; Liu, Y.-S.; and Zwicker, M. 2021. Neural-Pull: Learning Signed Distance Function from Point clouds by Learning to Pull Space onto Surface. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 7246–7257. PMLR.

Ma, B.; Zhou, J.; Liu, Y.-S.; and Han, Z. 2023. Towards better gradient consistency for neural signed distance functions via level set alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17724–17734.

Mescheder, L.; Oechsle, M.; Niemeyer, M.; Nowozin, S.; and Geiger, A. 2019. Occupancy Networks: Learning 3D Reconstruction in Function Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J.-M., eds., *Computer Vision – ECCV 2020*, 405–421. Cham: Springer International Publishing. ISBN 978-3-030-58452-8.

Oechsle, M.; Peng, S.; and Geiger, A. 2021. UNISURF: Unifying Neural Implicit Surfaces and Radiance Fields for Multi-View Reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 5589–5599.

Ren, Y.; Zhang, T.; Pollefeys, M.; Süsstrunk, S.; and Wang, F. 2023. Volrecon: Volume rendering of signed ray distance functions for generalizable multi-view reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16685–16695.

Schonberger, J. L.; and Frahm, J.-M. 2016. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4104–4113.

Schönberger, J. L.; Zheng, E.; Frahm, J.-M.; and Pollefeys, M. 2016. Pixelwise View Selection for Unstructured Multi-View Stereo. In Leibe, B.; Matas, J.; Sebe, N.; and Welling, M., eds., *Computer Vision – ECCV 2016*, 501–518. Cham: Springer International Publishing. ISBN 978-3-319-46487-9.

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

Wang, Y.; He, X.; Peng, S.; Lin, H.; Bao, H.; and Zhou, X. 2023. AutoRecon: Automated 3D Object Discovery and Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 21382–21391.

Xu, Q.; and Tao, W. 2019. Multi-Scale Geometric Consistency Guided Multi-View Stereo. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Yao, Y.; Luo, Z.; Li, S.; Fang, T.; and Quan, L. 2018. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (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. *Computer Vision and Pattern Recognition (CVPR)*.

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.

Yu, Z.; Peng, S.; Niemeyer, M.; Sattler, T.; and Geiger, A. 2022. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in neural information processing systems*, 35: 25018–25032.

Zhang, J.; Yao, Y.; Li, S.; Luo, Z.; and Fang, T. 2020. Visibility-aware Multi-view Stereo Network. *British Machine Vision Conference (BMVC)*.

Zhang, W.; Xing, R.; Zeng, Y.; Liu, Y.-S.; Shi, K.; and Han, Z. 2023. Fast Learning Radiance Fields by Shooting Much Fewer Rays. *IEEE Transactions on Image Processing*, 32: 2703–2718.

Zhou, J.; Ma, B.; Li, S.; Liu, Y.-S.; and Han, Z. 2023. Learning a More Continuous Zero Level Set in Unsigned Distance Fields through Level Set Projection. In *Proceedings of the IEEE/CVF international conference on computer vision*.

Zhou, J.; Ma, B.; Liu, Y.-S.; Fang, Y.; and Han, Z. 2022a. Learning consistency-aware unsigned distance functions progressively from raw point clouds. *Advances in Neural Information Processing Systems*, 35: 16481–16494.

Zhou, J.; Wen, X.; Ma, B.; Liu, Y.-S.; Gao, Y.; Fang, Y.; and Han, Z. 2022b. 3D-OAE: Occlusion Auto-Encoders for Self-Supervised Learning on Point Clouds. *arXiv preprint arXiv:2203.14084*.

Zhu, B.; Yang, Y.; Wang, X.; Zheng, Y.; and Guibas, L. 2023. VDN-NeRF: Resolving Shape-Radiance Ambiguity via View-Dependence Normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 35–45.