ICCV
#2711

SparseRecon: Neural Implicit Surface Reconstruction from Sparse Views with Feature and Depth Consistencies

Anonymous ICCV submission

Paper ID 2711

Abstract

Surface reconstruction from sparse views aims to reconstruct a 3D shape or scene from few RGB images. However, existing generalization-based methods do not generalize well on views that were unseen during training, while the reconstruction quality of overfitting-based methods is still limited by the limited geometry clues. To address this issue, we propose SparseRecon, a novel neural implicit reconstruction method for sparse views with volume rendering-based feature consistency and uncertainty-guided depth constraint. Firstly, we introduce a feature consistency loss across views to constrain the neural implicit field. This design alleviates the ambiguity caused by insufficient consistency information of views and ensures completeness and smoothness in the reconstruction results. Secondly, we employ an uncertainty-guided depth constraint to back up the feature consistency loss in areas with occlusion and insignificant features, which recovers geometry details for better reconstruction quality. Experimental results demonstrate that our method outperforms the state-of-the-art methods, which can produce high-quality geometry with sparse-view input, especially in the scenarios on small overlapping views.

1. Introduction

As one of the important tasks in computer vision, 3D reconstruction has attracted lots of research attentions in recent years. With the advancement of deep learning, 3D reconstruction using neural implicit representations based on point clouds [25, 32, 54, 55] or images [22, 35, 44, 56] becomes a popular research topic. Although existing methods [5, 35, 37, 40, 44, 52] that directly use images have made great progress in terms of the reconstruction quality and reconstruction speed, they require a large number of dense views as supervision. When the number of available views is limited, current reconstruction methods usually struggle to reconstruct high-quality surfaces.

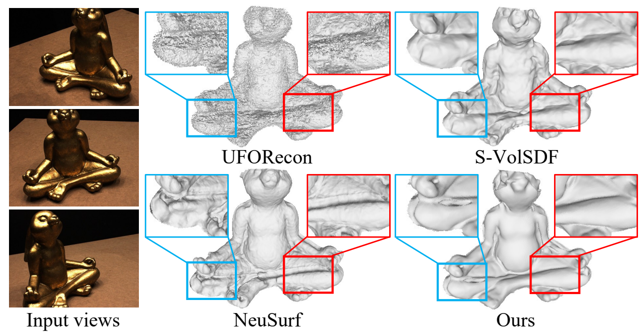

Figure 1. Given only 3 input images with large view angle change, our method can reconstruct a smoother surface compared to the state-of-the-art methods, such as UFORecon [27], S-VolSDF [39] and NeuSurf [13]. The details of each surface are shown in the colored boxes.

Existing methods for sparse view reconstruction can be mainly classified into two categories: generalization-based methods and overfitting-based methods. The generalization-based methods [21, 23, 27, 29, 30] emphasize the generalization of sparse-view reconstruction, but they are mainly effective in scenarios with large view overlaps. In cases with views that were unseen during training, the quality of the reconstructed surface degenerates significantly, as shown in Figure 1. Meanwhile, it takes a long time to pre-train these methods on large-scale data. Instead, overfitting-based methods [13, 14, 39, 45, 46] typically fit the 3D geometry directly from the sparse views by leveraging geometry clues. They show promising capability in reconstructing higher-quality geometric surfaces with small-overlapping views. However, the reconstruction quality of the existing methods is still unsatisfactory.

In this paper, we introduce a multi-view feature consistency loss based on volume rendering and an uncertainty-guided depth constraint to learn neural signed distance functions. This approach allows us to achieve high-quality mesh reconstruction on more challenging sparse views with small overlap.

For the *feature consistency loss*, we first employ the pre-trained Vis-MVSNet [49] to obtain depth features from the input images. Then, within a neural implicit rendering framework, the sampled 3D points along the rays emitted from the reference image are projected to the source image and the reference image. This allows us to acquire source features and reference features of each 3D point and measure the similarity between these two kinds of features. Finally, the feature similarity for each 3D point along the rays is accumulated through volume rendering, thus yielding the feature similarity associated with the rays. During optimization, we pursue higher feature similarity along the rays. Since the depth information is implicitly encoded with image features, feature consistency constraint can significantly alleviate the ambiguity issues arising from insufficient consistency of sparse views and low-texture during reconstruction.

For the *uncertainty-guided depth prior constraint*, we follow MonoSDF [46], utilizing a pre-trained network to acquire depth priors for each image, and then use it to constrain the regions with uncertain depth. However, monocular depth priors do not have consistent scales to the ground truth depth, which are hard to get calibrated to ground truth either due to the distortion. To effectively leverage the depth priors and provide proper supervision for occluded or under-constrained regions, we propose an uncertainty-guided depth prior constraint. First, we calibrate the depth priors using sparse point clouds obtained from COLMAP [31]. Then, during training, we compute the depth confidence from the rendered depth and impose the depth prior constraint only in regions with low confidence. This constraint helps infer more accurate geometry in occluded or under-constrained regions, minimizing the negative impact of depth prior errors on well-constrained regions.

We evaluate our methods on several widely used benchmarks and report the state-of-the-art results. In summary, our main contributions are as follows.

- We propose a novel feature consistency loss based on volume rendering. It can effectively constrains the neural radiance field by leveraging feature consistency among multiple views, improving the performance in sparse-view reconstruction tasks.
- By incorporating depth confidence, we utilize the calibrated depth prior more effectively to enhance geometric constraints, further improving the reconstruction quality.
- Extensive experiments on the well-known datasets, such as DTU [16] and BlendedMVS[43], demonstrate that our method outperforms existing sparse-view reconstruction methods and achieve the state-of-the-art results.

2. Related Work

2.1. Neural Implicit Reconstruction

Neural implicit reconstruction methods [5, 7, 20, 35–37, 40, 44, 46], have been rapidly developed based on neural volume rendering [26]. These methods introduce the Signed Distance Function (SDF) as the implicit representation of 3D surfaces in volume rendering, achieving multi-view 3D reconstruction. While these methods have made significant improvements in both reconstruction quality and speed, it is important to note that they heavily rely on multiple viewpoints during the optimization.

Generalization-based surface reconstruction with sparse views. In order to directly generalize the reconstruction results on sparse views, methods [21, 23, 27, 29, 30, 41] adopt the strategy of aggregating features from multiple view images to construct a feature volume, which is then used to predict the SDF for reconstructing the surface. VolRecon [30] uses transformers [17] to aggregate multi-view features, C2F2NeUS [41] employs cascade architecture to construct a volume pyramid, while ReTR [21] and UFORecon [27] aggregates multi-level features. These methods require pretraining on large-scale datasets, which typically takes several days. Moreover, when there is a significant domain gap between the testing and training data, they all fail to reconstruct shapes effectively.

Overfitting-based surface reconstruction with sparse views. In contrast, overfitting-based methods directly fit the 3D geometry from the sparse images by geometric prior constraints. MonoSDF [46] employs depth and normal priors to achieve sparse reconstruction with small-overlapping views. However, such priors come with errors, and it does not fully leverage inter-view consistency, resulting in lower reconstruction quality. S-VolSDF [39] employs probability volumes obtained from MVS [9] models to guide the rendering weight estimated by VolSDF [44]. This improves the reconstruction results in sparse views with small overlap. However, the uncertainties in volumes make negative impact on the reconstruction surface, leading to surface roughness or significant defects. More recently, NeuSurf [13] leverages sparse point clouds and employs CAP-UDF [54] to construct an implicit geometric prior to improve the reconstruction quality of sparse views. However, when the sparse point cloud fails to cover the majority of positions on the object surface, effective implicit geometric prior information cannot be obtained, which does not improve the reconstruction quality. In contrast, our method employs more robust feature priors, calculates feature consistency based on volume rendering, and simultaneously utilizes depth priors to optimize the occluded regions, ultimately resulting in high-quality geometric surface.

ICCV #2711

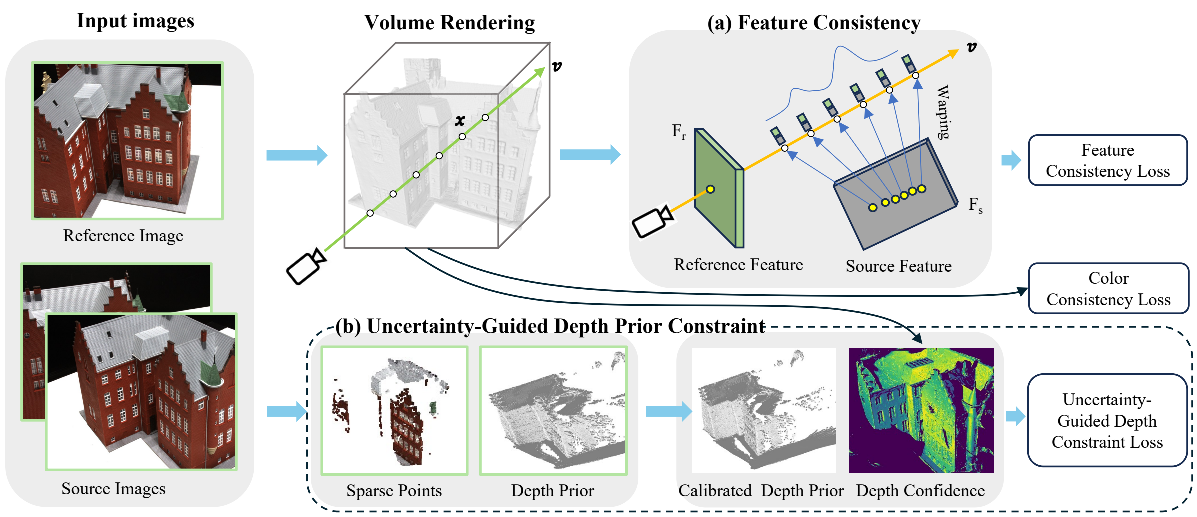

Figure 2. SparseRecon consists of two main parts. (a) Volume rendering-based feature consistency constraint. We extract features from the reference image and source images. For a ray emitted from the reference image, we project each sampled point on the ray onto the source images to obtain the corresponding features. Then, the volume rendering-based feature consistency loss is calculated using the corresponding features on the reference image. (b) Uncertainty-guided depth prior constraint. We use another pre-trained network to obtain the depth prior of the reference image and calibrate it with the sparse point cloud obtained by COLMAP. Then, we calculate the confidence of the rendered depth, so that the calibrated depth prior only constrains areas with low confidence.

2.2. Gaussian Splatting.

Gaussian Splatting [18] has achieved unprecedented optimization speed and rendering quality in the task of novel view synthesis. However, since the Gaussians are unorganized, the discrete and unstructured points make it difficult to extract 3D surfaces through post-processing. To address this issue, some methods introduce regularization terms [10], convert 3D Gaussians to 2D surfels [4, 12], acquire opacity fields through rays [47], improve the depth rendering algorithm [1] of 3DGS, or jointly optimize 3DGS with neural radiance fields [2, 24, 53]. However, these methods are only applicable to dense views. Recently, FatesGS [14] achieves fast sparse-view reconstruction by leveraging depth priors and on-surface feature consistency constraints. However, due to the poor convergence of the on-surface feature consistency constraint and the inaccuracy of the depth priors, the reconstruction results still exhibit roughness or noticeable defects.

2.3. Sparse View Synthesis.

In addition, the novel view synthesis from sparse views is another category of work closely related to sparse view reconstruction. Depending on the technical framework, these works can be categorized into NeRF-based methods [6, 15, 28, 33, 34, 42, 48] and Gaussian Splatting-based methods [3, 11, 19, 51, 57]. This line of research also employs a limited number of views as input. However, they solely focus on the rendering quality of novel views rather than surface reconstruction, which are not designed specifically for the accurate geometric surface reconstruction. Due to the discernible bias (i.e. inherent geometric errors) [35] caused by the conventional volume rendering method or inconsistencies in depth that appear in Gaussian rendering, current sparse view synthesis methods still fail to correctly reconstruct high-fidelity geometric surfaces.

3. Method

The overview of our method is depicted in Figure 2. We introduce a novel feature consistency loss and an uncertainty-guided depth constraint based on the NeuS [35] framework. In this section, we first explain how to compute feature consistency for sampled points along rays. Then we explain how to enhance geometric constraints using depth priors and depth uncertainty. Thirdly, we introduce the color consistency loss. Finally, we present the overall loss function for optimization.

3.1. Volume Rendering-based Feature Consistency

First, we use a pre-trained MVS network [49] to extract the features from both the reference image and the source image. Given a ray emitted from the reference image, let $p_r(0)$ denote the point where a ray intersects the reference image. And for each point x_i along the ray, we denote its projection on the source image as $p_s(i)$. Then, we bilinearly interpolate $F_r(0)$ and $F_s(i)$ at points $p_r(0)$ and $p_s(i)$ on image features, respectively. Formally, we define the feature con-

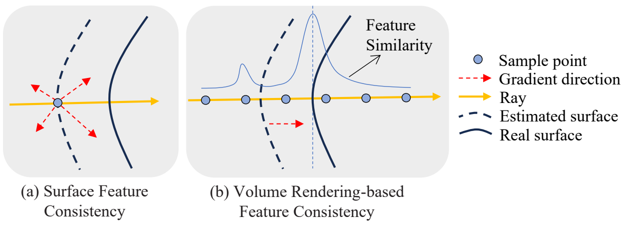

Figure 3. The illustration of (a) on-surface feature consistency and (b) feature consistency with volume rendering.

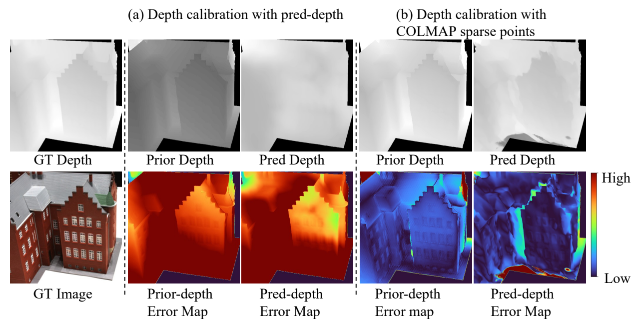

Figure 4. The illustration of predicted depth produced by different depth prior utilization methods, along with the corresponding error maps. (a) Calibrate the depth prior using the predicted depth during training. (b) Calibrate the depth prior using the COLMAP sparse point cloud.

sistency loss function as follows,

$$L_{feat} = M^{occ}(1 - \frac{1}{N}\sum_{i=1}^{N} w_i f_{cos}(F_r(p_r(0)), F_s(p_s(i))), \tag{1}$$

where f_{cos} is the cosine similarity, and w_i corresponds to the weight for each point along the ray. $p_s(i) = K(Rx_i + t)$ is the projection of x_i in source view, and $[K; R; t]$ is the camera parameters of source view. M^{occ} is the occlusion mask.

Although MVSDF [50], NeuSurf [13] and FatesGS [14] also employ feature consistency constraints, they just leverage the intersection point between a camera ray and the object's surface. Then, this intersection point gets projected onto adjacent views to obtain the corresponding image features for the purpose of comparing features at this point across multiple views. In sparse view scenarios, the estimated positions of surface points can easily deviate significantly, making the on-surface feature consistency loss not converge. NeuSurf [13] and FatesGS [14] utilize sparse point clouds generated by COLMAP [31] as priors, enabling it to obtain partially accurate positions of surface points, thereby allowing the on-surface feature consistency loss to be more effectively leveraged. However, in regions of lacking surface points, the on-surface feature consistency loss cannot ensure the attainment of high-quality geometric surfaces.

Figure 3 illustrates the difference between on-surface feature consistency and volume rendering-based feature consistency. Due to the uncertainty of gradient direction, the constraint solely relying on surface point features is challenging to be optimized. In contrast, our method does not require the prior estimation of surface points, it calculates feature consistency on all sampling points along the ray, and provides more reasonable and comprehensive supervision to the implicit field, thereby addressing the convergence issue that may arise in sparse reconstruction for MVSDF [50] and NeuSurf [13].

3.2. Uncertainty-Guided Depth Constraint

Although multi-view features offer more robust constraint than image colors, they are ineffective for occluded regions. Due to the limited number of views, some regions may only be visible from a single viewpoint. To enhance geometric constraints, we employ depth priors to supervise the radiance field. However, monocular depth priors are not perfect and accurate. Although MonoSDF [46] has already takes the inaccuracy of depth priors into account, i.e., it aligns depth priors using rendered depth during training. However, the rendered depth during training is inaccurate, resulting in significant errors in the calibrated depth priors. This ultimately leads to the accumulation of errors during training, which results in inaccurate reconstructions. Figure 4 (a) shows the calibrated depth prior and rendered depth obtained by MonoSDF [46], as well as their error maps compared to the ground truth depth. It can be seen that both the calibrated depth prior and the rendered depth are with large errors. Therefore, MonoSDF [46] uses a weight annealing stategy to anneal the weight of depth loss to 0 during the first 200 training epochs.

Another trivial approach is to calibrate the depth priors using the sparse point cloud obtained from COLMAP [31]. Since the sparse points are generally located on the geometric surface of the object, their depth is relatively accurate. Therefore, calibrating the depth priors using the sparse point cloud can lead to more accurate depth priors. Figure 4 (b) shows the depth priors calibrated with the sparse point cloud, and the depth rendered with the depth priors as a constraint, as well as their error maps compared to the ground truth depth. It indicates that the depth priors calibrated to the point cloud from the COLMAP [31] are more accurate. Therefore, we can use them as an constraint leads to more precise rendered depth.

However, due to the distortions in monocular depth priors, it is impossible to perfectly align them with the ground truth depth. Even after calibration, the depth priors still exhibit noticeable errors when compared to the ground truth depth. In sparse view scenarios, occlusions and insufficient

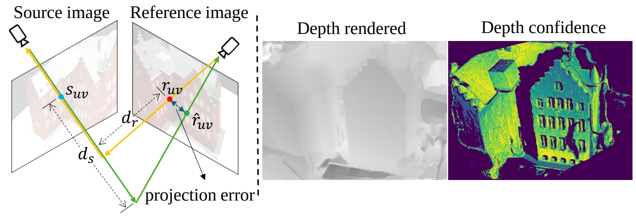

Figure 5. Left: the method of obtaining the confidence of rendered depth. Right: the rendered depth and the depth confidence.

constraints are more common, leading to significant discrepancies between the geometry of occluded regions and the real surface. Therefore, to achieve more accurate geometry in these under-constrained regions while avoiding the negative impact of depth prior errors on well-constrained regions, we propose an uncertainty-guided depth prior constraint method to more effectively utilize the depth priors. Specifically, we apply depth prior constraints in regions with depth uncertainty, while refraining from using them in regions with high depth confidence.

To obtain the confidence of the rendered depth, we employ a method to evaluate the multi-view depth projection consistency. As shown in Figure 5, for a specific pixel r_{uv} in the reference image with depth d_r, it can be mapped to a neighboring image through the homography matrix H_{rs}, leading to a pixel s_{uv},

$$s_{uv} = H_{rs} r_{uv}, \tag{2}$$

$$H_{rs} = M_s M_r^{-1}, \tag{3}$$

where M_r and M_s are the projection matrices corresponding to the reference and source views, respectively. Similarly, we can map the pixel s_{uv} in the source view to the reference view using the projection matrix H_{sr} and its corresponding depth d_s, resulting in $\hat{r}_{uv}$. The forward and backward projection distance error reflect the accuracy of depth predictions, so we take it as the depth confidence, which is defined as

$$C_d = \begin{cases} \frac{1}{e^{\|r_{uv} - \hat{r}_{uv}\|}}, & \text{if } \|r_{uv} - \hat{r}_{uv}\| \leqslant 1 \\ 0, & \text{if } \|r_{uv} - \hat{r}_{uv}\| > 1 \end{cases} \tag{4}$$

The right side of Figure 5 shows the rendered depth and the corresponding depth confidence.

Correspondingly, the depth uncertainty is defined as $U_d = 1 - C_d$. Meanwhile, we can set a threshold τ for depth confidence C_d to obtain the occlusion mask $M^{occ} = \{C_d > \tau\}$.

For depth calibration, we leverage COLMAP [31] to obtain a sparse point cloud $\{X : x_1, x_2 \ldots x_i \in R\}$ and visibility flags indicating which keypoints are visible from view I. Given the camera parameters P of view I, we estimate the depth $\bar{D}_i$ of keypoints by computing the distance from the visible keypoints x_i to the camera center o. Then, we calibrate the monocular depth prior $\hat{D}$ with $\bar{D}_i$, it can be defined as $\bar{D} \approx a\hat{D} + b$, where a is the scale factor and b is the shift factor, obtained through the least squares method. Formally, the depth constraint loss is defined as,

$$L_{depth} = \sum_{r \in R} U_d \left\| (a\hat{D} + b) - D_{pred} \right\|^2. \tag{5}$$

3.3. Color Consistency Constraint

Although feature consistency constraint can ensure that the reconstruction does not suffer from severe artifacts, it does not provide sufficient supervision to reconstruct fine geometric details. Conversely, in cases with rich textures, image color constraint can refine the geometric details. Therefore, following the NeuralWarp [5], pixel warping loss and patch warping loss are used in our method as multi-view color consistency loss functions,

$$\begin{aligned} L_{color} = &\sum_{r \in R} M^{occ} d_{pixel}(C(r), C_s(r)) \\ &+ \sum_{r \in R} M^{occ} d_{patch}(P(r), P_s(r)), \end{aligned} \tag{6}$$

where $C(r)$ and $C_s(r)$ are the ground truth color of the pixel from which the ray emits and the rendered color, respectively, $P(r)$ and $P_s(r)$ are the ground truth color of the patch corresponding to the ray and the rendered patch color, respectively. d_{pixel} is the loss metric for pixel color, where we use $L1$ loss as d_{pixel}. d_{patch} is the loss metric for patch color, where we use the Structural Similarity Index Measure (SSIM [38]) as d_{patch}.

3.4. Training Loss

In addition to the above-mentioned three loss functions, we also use the Eikonal loss [8] used in NeuS [35]. We define the overall loss function as follows:

$$L = L_{feat} + \alpha L_{depth} + L_{color} + \beta L_{eik}, \tag{7}$$

L_{eki} is the Eikonal loss [8], used to regularize the SDF values of sampled points, defined as

$$L_{eki.} = \frac{1}{mn} \sum_{i,k} (\|\nabla f(x_{i,k})\|_2 - 1)^2. \tag{8}$$

4. Experiments

4.1. Dataset

We evaluate our method on DTU [16] and BlendedMVS [43] dataset. For the DTU [16] dataset, to avoid using the scenes that have already been used as training data on the pretrained Vis-MVSNet [49] model, we select the same 11

ICCV
#2711

Methods	21	24	34	37	38	40	82	106	110	114	118	Mean CD ↓
VolSDF [44]	5.47	4.38	3.15	7.38	1.88	6.70	5.19	4.67	2.79	1.32	1.83	4.07
NeuS [35]	5.63	3.58	6.00	4.60	2.57	4.53	1.91	4.18	5.46	1.19	4.16	3.98
NeuralWarp [5]	2.53	1.88	0.74	1.80	**0.84**	11.50	2.64	2.10	4.37	1.19	2.63	2.93
MonoSDF [46]	4.14	5.92	1.39	4.55	2.19	2.14	2.36	5.62	4.58	1.63	3.02	3.41
Vis-MVSNet [49]	3.39	4.44	0.85	3.36	1.69	3.35	3.35	2.34	2.16	0.74	1.83	2.50
MVSDF [50]	4.31	4.71	1.65	6.37	1.77	4.47	3.61	1.87	1.67	1.25	1.69	3.03
2DGS [12]	4.47	3.54	3.48	4.13	4.25	3.61	4.83	2.40	2.97	1.35	2.17	3.38
PGSR [1]	5.58	4.01	3.15	5.19	4.55	3.65	5.57	2.35	1.91	0.57	1.55	3.46
SparseNeuS$_{ft}$ [23]	3.48	4.37	2.92	4.76	2.79	3.73	2.80	1.86	3.10	1.15	2.29	3.02
VolRecon [30]	2.72	3.07	1.82	4.32	2.14	3.04	3.00	2.56	2.81	1.49	3.22	2.75
GenS$_{ft}$ [29]	5.86	7.67	3.62	8.57	5.37	5.41	5.48	6.04	5.29	4.69	4.35	5.67
ReTR [21]	2.67	3.37	1.62	3.68	1.87	3.40	3.67	2.84	2.85	1.56	2.35	2.72
UFORecon [27]	**1.84**	1.52	0.79	2.58	1.00	1.82	1.72	1.20	0.93	0.66	1.26	1.39
S-VolSDF [39]	2.45	3.08	1.33	3.09	1.22	3.21	1.91	1.51	1.23	0.74	1.2	1.91
SparseCraft [45]	2.88	2.42	0.92	2.97	1.58	2.78	2.51	1.10	5.24	0.65	0.88	2.16
NeuSurf [13]	7.60	1.43	2.93	3.18	1.53	2.86	1.86	1.09	1.41	**0.37**	**0.62**	2.26
FatesGS [14]	3.98	1.32	2.53	2.85	3.36	2.71	3.76	1.49	0.85	0.47	1.06	2.22
Ours	2.14	**1.26**	**0.72**	**1.46**	0.86	**1.39**	**1.37**	**0.94**	**0.77**	0.44	0.83	**1.11**

Table 1. Quantitative results of Chamfer Distance (CD↓) on DTU dataset with *3 small-overlapping* images. The methods are divided into three categories, from top to bottom: (1) dense-view reconstruction methods related to ours, (2) generalization-based sparse-view reconstruction methods, and (3) overfitting-based sparse-view reconstruction methods. the best results are in *bold*, the second best are *underlined*.

scenes as in S-VolSDF [39]. The image resolution is set to 1600×1200. Similar to the S-VolSDF [39] and NeuSurf [13] methods, we select the views 22, 25, and 28 for the more challenging reconstruction of small overlaps.

For the BlendedMVS [43] dataset, we follow the S-VolSDF [39] to use the same 9 challenging scenes, with 3 small-overlapping views for each scene. The image resolution is set to 768×576.

4.2. Implementation Details

We use the same network architecture and initialization strategy as NeuS [35] and incorporated our volume rendering feature consistency loss, uncertainty-guided depth constraint loss, and color consistency loss. For the weight factors in the loss functions Eq. 7, we set the α for the uncertainty-guided depth prior constraint loss L_{depth} to 0.5 and the β for the Eikonal loss L_{eik} to 0.1. Each scene is trained 100K iterations on a RTX3090 GPU. The patch warping term in the color consistency loss requires the surface point normals to calculate homographies, but the initial normals are too noisy [5], therefore, the patch warping loss is applied after 20k training steps. The threshold τ of the occlusion mask is set to 0.

4.3. Baseline

We compare our approach with three categories of methods. *Dense-view methods*: NeuS [35], VolSDF [39], NeuralWarp [5], Vis-MVSNet [49], MVSDF [50], 2DGS [12] and PGSR [1]. *Generalization-based methods*: SparseNeuS [23], VolRecon [30], GenS [29], ReTR [21] and UFORecon [27]. *Overfitting-based methods*: S-VolSDF [39], SparseCraft [45], NeuSurf [13] and FatesGS [14]. The reconstruction results for SparseNeuS [23] and GenS [29] are fine-tuned using 3 views for each scene.

4.4. Comparisons

Reconstruction on DTU. For a comprehensive comparison, we evaluate the baselines and our method on both small-overlapping and large-overlapping views. Following baselines [13, 14, 23], we report the Chamfer Distance (CD) between the reconstruction surfaces and the ground truth point clouds. The CD results with small overlapping views are shown in Table 1. The meshes reconstructed by several methods using 3 views with small overlapping are shown in Fig. 6. For the generalization-based sparse reconstruction methods, we only show the reconstruction results of the latest UFORecon [27], as the reconstruction quality of other methods is lower than that of UFORecon [27]. The experimental results show that our method significantly improves the mesh quality with small overlap views, compared to the state-of-the-art sparse-view reconstruction methods. The results of large overlapping views are presented in the supplementary materials.

As shown in Figure 6, when input sparse views with

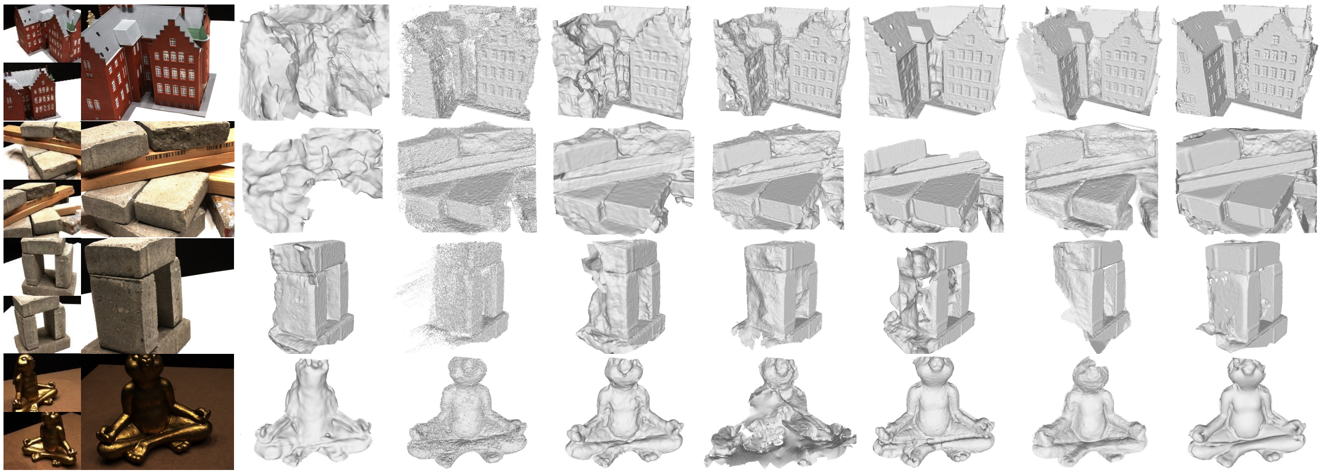

Figure 6. Visual comparison on DTU dataset with *3 small-overlapping* images.

small overlap, both MonoSDF [46] and SparseCraft [45] suffer from reconstruction ambiguity and failures, highlighting that relying solely on simplistic geometric prior constraints is insufficient to obtain complete and accurate meshes. UFORecon [27] shows significant roughness in its reconstruction results. S-VolSDF [39], NeuSurf [13] and FatesGS [14] exhibit noticeable reconstruction defects. Experimental results demonstrate that our method is effective in alleviating geometric and appearance ambiguities during the optimization process. This significantly enhances the quality of mesh reconstruction, especially in scenarios with small overlapping views and low texture.

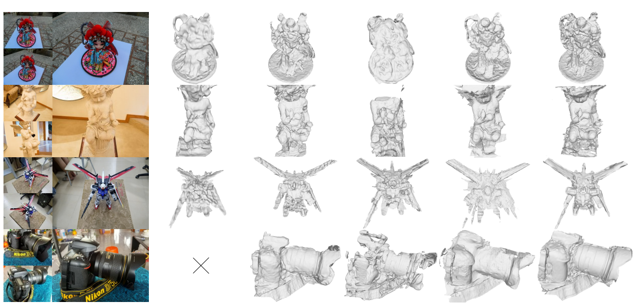

Figure 7. Visual comparison on BlendedMVS dataset. '×' indicates reconstruction failure.

Reconstruction on BlendedMVS. Figure 7 presents the visual comparison of reconstructed mesh for overfitting-based methods. With only 3 small-overlapping views provided, all of the generalization-based methods completely fail to reconstruct in the sparse setting of BlendedMVS[43] dataset, even if SparseNeuS [23] is fine-tuned. Therefore, the reconstruction results of these methods are not included in Figure 7. Compared to other methods, our approach can generate more complete and detailed meshes. Similarly, MonoSDF [46] fails to reconstruct either. The meshes generated by S-VolSDF [39], NeuSurf [13] and FatesGS [14] exhibit significant defects. Both NeurSurf [13] and FatesGS [14] use on-surface feature consistency constraints, but the reconstruction results are still not good enough. In contrast, our method achieves more comprehensive geometry and finer details by employing volume rendering-based feature consistency constraints. This highlights the advantages of our approach in geometric consistency. More visualizations are presented in the supplementary materials.

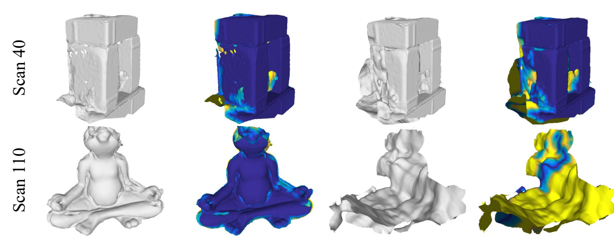

Figure 8. Reconstructed meshes and error maps on DTU dataset with different feature consistency losses.

4.5. Ablation Study

We evaluate the components of our method with 3 small-overlapping views by an ablation study on the DTU [16] dataset. To compare the depth loss L_{depth}^{mono} calibrated by rendered depth in MonoSDF [46] with our depth loss L_{depth}, we replace L_{depth} with L_{depth}^{mono} to evaluate it in our method. We also compare the volume rendering-based

Method	L_{color}	L_{feat}	L_{depth}	L_{depth}^{mono}	L_{feat}^{L1}	L_{feat}^{L2}	L_{feat}^{surf}	CD↓
Baseline								3.35
	✓							1.76
	✓	✓						1.47
	✓		✓					1.62
	✓	✓	✓					**1.11**
	✓	✓		✓				1.59
	✓		✓		✓			2.36
	✓		✓			✓		1.81
	✓		✓				✓	2.93

Table 2. Ablation studies on DTU dataset with 3 small-overlapping images.

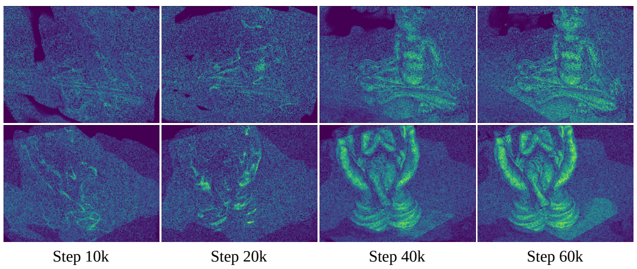

Figure 9. The variation of weighted feature similarity during training, brighter colors indicate higher feature similarity.

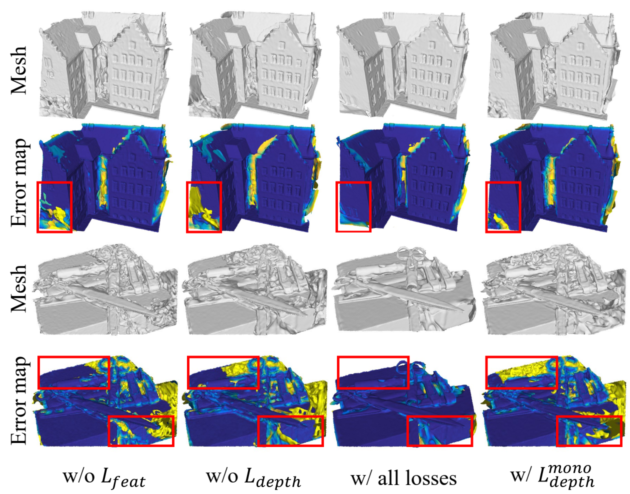

Figure 10. Visualization of reconstruction and error maps for scene scan24 and scan37 in DTU dataset with different losses. The differences of error maps are highlighted.

feature consistency loss calculated using L1 distance (denoted as L_{feat}^{L1}) and L2 distance (denoted as L_{feat}^{L2}) with our method using feature similarity distance. We found that feature similarity distance is better than both L1 and L2 distance, as shown in Table 2.

In addition, we replace our volume rendering-based feature consistency loss L_{feat} with the on-surface feature consistency loss L_{feat}^{surf} used in MVSDF [50] to compare the effects of two different loss functions. Figure 8 illustrates the

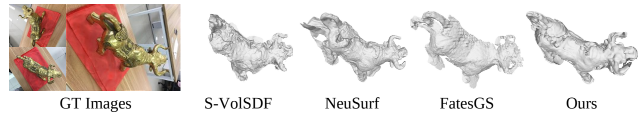

Figure 11. Failure case. For specular objects, the ambiguity in the color consistency constraint may lead to a rough surface.

reconstruction results and error maps on the DTU dataset when using different feature consistency losses, under the on-surface feature consistency loss L_{feat}^{surf}, the meshes show large artifacts.

Table 2 shows the average Chamfer Distance over all 11 scenes on DTU dataset using different losses. The experimental results indicate that both feature consistency loss and uncertainty-guided depth constraint improve the surface reconstruction.

Figure 9 illustrates the variation of the weighted feature similarity map during the training process. Brighter colors indicate higher feature similarity, demonstrating that our volume rendering-based feature consistency loss can provide effective constraints.

Figure 10 shows the reconstructed meshes and error maps for scene scan24 and scan37 on the DTU [16] dataset when using different losses. It can be observed that the mesh deteriorates with out the volume rendering-based feature consistency loss or the uncertainty-guided depth constraint loss, and the reconstruction quality drops when using the depth loss L_{depth}^{mono} in MonoSDF [46].

5. Conclusions

We propose a novel method for learning implicit representations from sparse views with small overlaps. Our novelty lies in a novel volume rendering-based feature consistency loss and an uncertainty-guided depth constraint. Extensive experiments on the DTU [16] and BlendedMVS [43] datasets show that our method surpasses existing state-of-the-art sparse-view reconstruction methods in terms of reconstruction quality.

Limitations. Although our method shows significant improvement over other sparse view reconstruction methods, there are still some limitations. Firstly, for specular objects, the ambiguity in the color consistency constraint may lead to a rough surface, as shown in Figure 11. Secondly, Following previous studies [13, 23, 39], the camera poses of sparse views are obtained from the training dataset. However, in some cases, it may not be possible to obtain accurate camera poses using SfM methods like COLMAP [31] due to the lack of texture in the images or excessive viewing angles. Additionally, the feature consistency constraint method requires a pre-trained network to extract image features. The accuracy of the features determines the performance of the feature consistency constraint.

References

[1] Danpeng Chen, Hai Li, Weicai Ye, Yifan Wang, Weijian Xie, Shangjin Zhai, Nan Wang, Haomin Liu, Hujun Bao, and Guofeng Zhang. PGSR: Planar-based gaussian splatting for efficient and high-fidelity surface reconstruction. *IEEE Transactions on Visualization and Computer Graphics*, PP: 1–12, 2024. 3, 6

[2] Hanlin Chen, Chen Li, and Gim Hee Lee. NeuSG: Neural implicit surface reconstruction with 3D gaussian splatting guidance. *arXiv preprint arXiv:2312.00846*, 2023. 3

[3] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. MVSplat: Efficient 3D gaussian splatting from sparse multi-view images. In *European Conference on Computer Vision*, pages 370–386. Springer, 2025. 3

[4] Pinxuan Dai, Jiamin Xu, Wenxiang Xie, Xinguo Liu, Huamin Wang, and Weiwei Xu. High-quality surface reconstruction using gaussian surfels. In *ACM SIGGRAPH 2024 Conference Proceedings*, 2024. 3

[5] François Darmon, Bénédicte Bascle, Jean-Clément Devaux, Pascal Monasse, and Mathieu Aubry. Improving neural implicit surfaces geometry with patch warping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6260–6269, 2022. 1, 2, 5, 6

[6] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12882–12891, 2022. 3

[7] 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. 2

[8] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *Proceedings of Machine Learning and Systems 2020*, pages 3569–3579. 2020. 5

[9] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2495–2504, 2020. 2

[10] Antoine Guédon and Vincent Lepetit. SuGaR: Surface-aligned gaussian splatting for efficient 3D mesh reconstruction and high-quality mesh rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5354–5363, 2024. 3

[11] Liang Han, Junsheng Zhou, Yu-Shen Liu, and Zhizhong Han. Binocular-guided 3D gaussian splatting with view consistency for sparse view synthesis. In *Advances in Neural Information Processing Systems*, 2024. 3

[12] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2D gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 Conference Proceedings*, 2024. 3, 6

[13] Han Huang, Yulun Wu, Junsheng Zhou, Ge Gao, Ming Gu, and Yu-Shen Liu. NeuSurf: On-surface priors for neural surface reconstruction from sparse input views. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2312–2320, 2024. 1, 2, 4, 6, 7, 8

[14] Han Huang, Yulun Wu, Chao Deng, Ge Gao, Ming Gu, and Yu-Shen Liu. FatesGS: Fast and accurate sparse-view surface reconstruction using gaussian splatting with depth-feature consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. 1, 3, 4, 6, 7

[15] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting NeRF on a Diet: Semantically consistent few-shot view synthesis implementation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5885–5894, 2021. 3

[16] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 406–413, 2014. 2, 5, 7, 8

[17] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning (ICML)*, pages 5156–5165, 2020. 2

[18] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4):1–14, 2023. 3

[19] Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Xin Ning, Jun Zhou, and Lin Gu. DNGaussian: Optimizing sparse-view 3D gaussian radiance fields with global-local depth normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20775–20785, 2024. 3

[20] Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. Neuralangelo: High-fidelity neural surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8456–8465, 2023. 2

[21] Yixun Liang, Hao He, and Yingcong Chen. ReTR: Modeling rendering via transformer for generalizable neural surface reconstruction. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 6

[22] Xinqi Liu, Jituo Li, and Guodong Lu. Reconstructing complex shaped clothing from a single image with feature stable unsigned distance fields. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 1

[23] Xiaoxiao Long, Cheng Lin, Peng Wang, Taku Komura, and Wenping Wang. SparseNeuS: Fast generalizable neural surface reconstruction from sparse views. In *European Conference on Computer Vision*, pages 210–227. Springer, 2022. 1, 2, 6, 7, 8

[24] 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. *arXiv preprint arXiv:2404.00409*, 2024. 3

[25] Baorui Ma, Yu-Shen Liu, and Zhizhong Han. Learning signed distance functions from noisy 3D point clouds via noise to noise mapping. In *International Conference on Machine Learning (ICML)*, 2023. 1

[26] 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

[27] Youngju Na, Woo Jae Kim, Kyu Beom Han, Suhyeon Ha, and Sung-Eui Yoon. UFORecon: Generalizable sparse-view surface reconstruction from arbitrary and unfavorable sets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5094–5104, 2024. 1, 2, 6, 7

[28] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. RegNeRF: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5480–5490, 2022. 3

[29] Rui Peng, Xiaodong Gu, Luyang Tang, Shihe Shen, Fanqi Yu, and Ronggang Wang. GenS: Generalizable neural surface reconstruction from multi-view images. In *Advances in Neural Information Processing Systems*, pages 56932–56945, 2023. 1, 2, 6

[30] Yufan Ren, Tong Zhang, Marc Pollefeys, Sabine Süsstrunk, and Fangjinhua Wang. 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*, pages 16685–16695, 2023. 1, 2, 6

[31] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, 2016. 2, 4, 5, 8

[32] Hui Tian, Chenyang Zhu, Yifei Shi, and Kai Xu. SuperUDF: Self-supervised udf estimation for surface reconstruction. *IEEE Transactions on Visualization and Computer Graphics*, 2023. 1

[33] Prune Truong, Marie-Julie Rakotosaona, Fabian Manhardt, and Federico Tombari. SPARF: Neural radiance fields from sparse and noisy poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4190–4200, 2023. 3

[34] Guangcong Wang, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. SparseNeRF: Distilling depth ranking for few-shot novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9065–9076, 2023. 3

[35] 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*, 2021. 1, 2, 3, 5, 6

[36] Yiqun Wang, Ivan Skorokhodov, and Peter Wonka. HF-NeuS: Improved surface reconstruction using high-frequency details. *Advances in Neural Information Processing Systems*, 35:1966–1978, 2022.

[37] Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. NeuS2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3295–3306, 2023. 1, 2

[38] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 5

[39] Haoyu Wu, Alexandros Graikos, and Dimitris Samaras. S-VolSDF: Sparse multi-view stereo regularization of neural implicit surfaces. *International Conference on Computer Vision*, 2023. 1, 2, 6, 7, 8

[40] Tong Wu, Jiaqi Wang, Xingang Pan, Xudong Xu, Christian Theobalt, Ziwei Liu, and Dahua Lin. Voxurf: Voxel-based efficient and accurate neural surface reconstruction. *International Conference on Learning Representations*, 2022. 1, 2

[41] Luoyuan Xu, Tao Guan, Yuesong Wang, Wenkai Liu, Zhaojie Zeng, Junle Wang, and Wei Yang. C2F2NeUS: Cascade cost frustum fusion for high fidelity and generalizable neural surface reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18291–18301, 2023. 2

[42] Jiawei Yang, Marco Pavone, and Yue Wang. FreeNeRF: Improving few-shot neural rendering with free frequency regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8254–8263, 2023. 3

[43] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. BlendedMVS: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1790–1799, 2020. 2, 5, 6, 7, 8

[44] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021. 1, 2, 6

[45] Mae Younes, Amine Ouasfi, and Adnane Boukhayma. SparseCraft: Few-shot neural reconstruction through stereopsis guided geometric linearization. In *European Conference on Computer Vision*, pages 37–56. Springer, 2024. 1, 6, 7

[46] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. MonoSDF: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in Neural Information Processing Systems*, 35:25018–25032, 2022. 1, 2, 4, 6, 7, 8

[47] Zehao Yu, Torsten Sattler, and Andreas Geiger. Gaussian Opacity Fields: Efficient adaptive surface reconstruction in unbounded scenes. *ACM Transactions on Graphics*, 43(6): 1–13, 2024. 3

[48] Yu-Jie Yuan, Yu-Kun Lai, Yi-Hua Huang, Leif Kobbelt, and Lin Gao. Neural radiance fields from sparse rgb-d images for

high-quality view synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8713–8728, 2022. 3

[49] Jingyang Zhang, Yao Yao, Shiwei Li, Zixin Luo, and Tian Fang. Visibility-aware multi-view stereo network. *The British Machine Vision Conference*, 2020. 2, 3, 5, 6

[50] Jingyang Zhang, Yao Yao, and Long Quan. Learning signed distance field for multi-view surface reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6525–6534, 2021. 4, 6, 8

[51] Jiawei Zhang, Jiahe Li, Xiaohan Yu, Lei Huang, Lin Gu, Jin Zheng, and Xiao Bai. CoR-GS: sparse-view 3d gaussian splatting via co-regularization. In *European Conference on Computer Vision*, pages 335–352. Springer, 2024. 3

[52] Wenyuan Zhang, Ruofan Xing, Yunfan Zeng, Yu-Shen Liu, Kanle Shi, and Zhizhong Han. Fast learning radiance fields by shooting much fewer rays. *IEEE Transactions on Image Processing*, 2023. 1

[53] Wenyuan Zhang, Yu-Shen Liu, and Zhizhong Han. Neural signed distance function inference through splatting 3d gaussians pulled on zero-level set. In *Advances in Neural Information Processing Systems*, 2024. 3

[54] Junsheng Zhou, Baorui Ma, Shujuan Li, Yu-Shen Liu, Yi Fang, and Zhizhong Han. CAP-UDF: Learning unsigned distance functions progressively from raw point clouds with consistency-aware field optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1, 2

[55] Junsheng Zhou, Baorui Ma, and Liu Yu-Shen. Fast learning of signed distance functions from noisy point clouds via noise to noise mapping. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1

[56] Tiansong Zhou, Jing Huang, Tao Yu, Ruizhi Shao, and Kun Li. HDhuman: High-quality human novel-view rendering from sparse views. *IEEE Transactions on Visualization and Computer Graphics*, 2023. 1

[57] Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. FSGS: Real-time few-shot view synthesis using gaussian splatting. In *European Conference on Computer Vision*, pages 145–163. Springer, 2025. 3