S-VolSDF: Sparse Multi-View Stereo Regularization of Neural Implicit Surfaces

Haoyu Wu Alexandros Graikos Dimitris Samaras
Stony Brook University
{haoyuwu,agraikos,samaras}@cs.stonybrook.edu

Abstract

Neural rendering of implicit surfaces performs well in 3D vision applications. However, it requires dense input views as supervision. When only sparse input images are available, output quality drops significantly due to the shape-radiance ambiguity problem. We note that this ambiguity can be constrained when a 3D point is visible in multiple views, as is the case in multi-view stereo (MVS). We thus propose to regularize neural rendering optimization with an MVS solution. The use of an MVS probability volume and a generalized cross entropy loss leads to a noise-tolerant optimization process. In addition, neural rendering provides global consistency constraints that guide the MVS depth hypothesis sampling and thus improves MVS performance. Given only three sparse input views, experiments show that our method not only outperforms generic neural rendering models by a large margin but also significantly increases the reconstruction quality of MVS models.

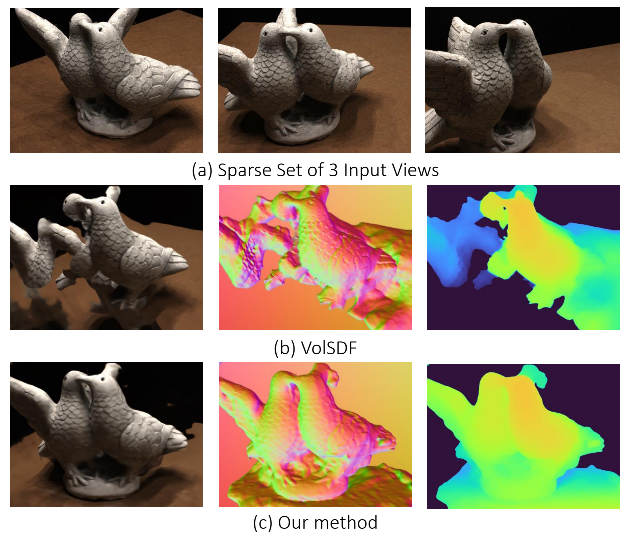


Figure 1. Shape-Radiance Ambiguity. In the last two rows, we compare the novel view synthesis results from VolSDF [74] and our model: RGB renderings (left), predicted normal maps (middle), and expected depth maps (right).

1. Introduction

Neural surface reconstruction techniques, coupled with coordinate-based neural network models, have become increasingly popular in the field of 3D vision [74, 61, 75]. Although these methods perform very well, they require dense input views as supervision. This is limiting for many real-world applications where sparse input images are the only source of information, such as robotics, augmented reality, autonomous driving, and scene reconstruction in-the-wild. As shown in Fig. 1, the reconstruction quality of a scene using VolSDF [74] (a state-of-the-art technique) drops significantly when only 3 views are used. This is due to the *shape-radiance ambiguity* problem [82].

The *shape-radiance ambiguity* [82] means that there is a high probability an incorrect geometry reconstruction satisfies the photometric constraint when it is visible from a single view only, as is in the case of sparse views. In that scenario, the photometric loss alone can not guide the model toward a correct solution. To regularize this, we need to constrain surface points to be visible from multiple views, hence, we need correspondences, as in multi-view stereo (MVS) [7, 38, 68, 71, 72, 9, 77, 21, 70, 80, 66, 69, 63, 84, 86, 16]. Thus, we propose to guide neural rendering optimization with information from MVS. The challenge is how to effectively incorporate the noisy MVS predictions into the neural rendering pipeline.

Many modern MVS methods [71, 21, 16, 9] integrate the evidence for each possible 3D point into a probability volume and regress depth from it. In order to avoid possible errors in MVS 3D point reconstruction, we do not use point estimates, but the whole probability volume. We also note that the rendering weights in neural rendering methods and the probability volume in MVS actually have the same meaning: the probability that a point at a particular location is visible by multiple views. Based on recent MVS literature [46], we can think of all possible 3D points on a ray as interior or exterior to the object (i.e. a binary classification problem). Thus, we can treat the MVS probability volume as a set of noisy labels for the rendering weights (i.e. occupancy values). Posing neural rendering as a classification problem allows the use of cross entropy loss to optimize neural rendering methods. However, as shown by the classification literature [85, 54], the cross entropy loss is sensitive to noisy labels. Instead, we adopt a generalized cross entropy loss [85] to reduce the penalty on false positive MVS predictions and thus increase the optimization's tolerance to noise.

In order to produce our final geometry, we want to take advantage of global consistency constraints including photometric consistency and surface smoothness imposed by neural rendering. Thus, we propose to incorporate neural surface reconstruction into coarse-to-fine MVS models. Specifically, we use the coarse stage MVS predictions to regularize neural surface optimization. Then, we use the rendered depth maps to guide the next stage's depth hypothesis sampling in MVS. Moreover, neural surface optimization only requires 10-15 minutes in current hardware to obtain good results because of strong geometry cues from MVS. As a result, we obtain much better surface reconstruction than either MVS or neural rendering alone, at a relatively fast speed.

In this paper, we propose *S-VolSDF*, a novel approach that leverages multi-view stereo priors to optimize neural surface reconstruction with sparse input views. Our main contributions are as follows:

- We propose a simple but effective noise-tolerant cost function that combines multi-view stereo with neural volumetric surface reconstruction methods, so their optimization is regularized by the probability volumes of MVS methods.
- We integrate neural surface reconstruction into multiple coarse-to-fine MVS models. Our method consistently improves depth estimation for better MVS performance at a faster speed.
- We evaluate our method on surface reconstruction and novel view synthesis on the DTU [1] and BlendedMVS [73] datasets. Our reconstruction is significantly better than both neural rendering and MVS models.

2. Related Work

2.1. Multi-View Stereo (MVS)

Traditional multi-view stereo uses representations such as depth maps, point clouds, and volumetric representations [18]. Depth map based methods [5, 20, 56, 50, 67] typically rely on a reference image and additional nearby source images for depth estimation. Point cloud based methods [19, 32, 35] attempt to optimize a collection of patches that best describe a 3D scene. Volumetric methods [29, 27, 31, 51, 11, 58, 22, 52, 17, 79] often aggregate information into a global representation such as a volume or mesh.

Deep-learning MVS methods [7, 38, 68, 71, 72, 9, 77, 21, 70, 80, 66, 69, 63, 84, 86, 16] typically use depth maps as 3D representations and follow the steps below: i) they use a differentiable homography to aggregate features from nearby views and build the cost volume, ii) they use a 3D CNN to regularize the cost volume and regress the depth and finally, iii) by applying a *softmax* function, they obtain a probability volume from the cost volume. A winner-takes-all technique is often used to determine the depth. Cascade cost volumes [9, 21, 70, 80, 16] and recurrent cost volume regularization [66, 69, 72] further reduce memory consumption. The cascade cost volume is constructed in a coarse-to-fine manner that first regresses a coarse depth in low resolution and then predicts finer depth values in higher resolution based on the depth range inferred from the coarse result.

MVS explicitly forces surface points to be visible from multiple views. This property prevents degenerate geometry in the case of sparse input views. However, the correspondence problem is often hard to solve, which introduces significant noise in the predicted geometry. Furthermore, the use of the *argmax* operation (winner-takes-all) removes potentially correct predictions in the MVS probability volume and introduces further noise. Thus, we propose to directly use information from the probability volume instead of the noise-prone MVS point estimates.

2.2. Neural Volumetric Representations

Neural volumetric representations are popular in 3D reconstruction [24, 75, 43, 28, 33, 61, 74, 44, 81, 12] and novel view synthesis [53, 36, 41, 48, 57, 2, 39, 45, 75]. NeRF [41], the most well-known method, is based on the volume rendering equation [40, 26] and stores 3D information inside a neural network in the form of a compact Multi-layer Perceptron (MLP). Due to the expressive power of the neural network, it is able to model high-quality details and reconstruct complex 3D structures with a relatively small storage cost. VolSDF builds on NeRF with improved volumetric rendering of implicit surfaces. However, as shown in Fig. 1, in the case of sparse input views, the quality of the VolSDF reconstruction drops significantly because of the radiance-ambiguity problem described in Sec. 1.

Regularization-based approaches are simple, but efficient ways to mitigate this problem, using priors such as smoothness [42, 44], cross-view semantic similarity constraint [23], normal priors [60], and depth priors [47, 65]. DS-NeRF [14] utilizes estimated depth from structure-from-motion [49]. MonoSDF [78] and SparseNeRF [59] utilize monocular depth estimation. Monocular depth estimation is often not accurate, only roughly approximating shapes, and may lead to sub-optimal results. Sensor depth [15, 30] and MVS [81, 87] have also been adopted to regularize the training of neural rendering models. Although MVS is a strong prior in general, it can be unreliable when the MVS prediction is noisy with sparse input views.

A different approach is to increase the generalization ability of neural rendering by utilizing priors derived from a larger model trained on multi-view image datasets [37, 6, 62, 76, 34, 10, 34, 55]. PixelNeRF [76] is conditioned on features extracted by a CNN. MVSNeRF [6] forms a neural volume from the cost volume obtained by warping image features, and is conditioned on this neural volume. IBRNet [62] aggregates features from nearby views to infer geometry and adopts an image-based rendering approach. GeoNeRF [25] utilizes a cascaded cost volume and an attention-based technique to aggregate information from different views. SparseNeuS [37] proposes cascaded geometry reasoning and consistency-aware fine-tuning. These methods considerably improve reconstruction, but our experiments show that their results still suffer from entanglement of texture with geometry, and inconsistencies between views.

Our method differs from generic neural rendering methods like MVSNeRF and GeoNeRF in that we explicitly utilize the MVS prior through noise-tolerant test-time optimization. In contrast, generic methods implicitly utilize the MVS prior by conditioning the rendering MLP on features derived from the cost volume, which may not work well in challenging sparse-input scenarios. In Sec. 4.3, we show our approach outperforms generic methods to effectively and reliably disentangle texture and geometry.

3. Method

Figure 2. Our proposed method improves the quality of depth maps obtained from the coarse stage multi-view stereo (MVS) by introducing noise-tolerant optimization techniques. The resulting depth maps then guide depth hypothesis sampling in the finer stage MVS, leading to more accurate and detailed 3D reconstructions.

We propose a novel way to integrate neural volume rendering with multi-view stereo algorithms. Specifically, we adopt VolSDF [74] for the neural surface reconstruction and notice that with sparse input views, VolSDF's reconstruction quality degrades dramatically. To mitigate this, we propose *S-VolSDF* that makes use of the correspondence-aware probability volume from MVS algorithms. Fig. 2 and Fig. 3 provide an overall illustration of our method.

3.1. Background

Volume Rendering of Implicit Surfaces. We use forward volume rendering [40, 26, 41] as our differentiable volumetric representation of the 3D scene and apply VolSDF [74]. VolSDF represents scene geometry as a signed distance function (SDF), which is subsequently transformed into density values for volume rendering. For each pixel, we sample points between the near and far depths along the ray $\mathbf{r}$ and approximate the pixel color $\hat{C}$ by:

$$\begin{aligned} \hat{C}(\mathbf{r}) &= \sum_{i=1}^{N} \mathbf{w}_i \cdot \mathbf{c}_i, \\ \text{where} \quad \mathbf{w}_i &= T_i \left(1 - \exp\left(-\sigma_i \delta_i\right)\right), \\ T_i &= \exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right). \end{aligned} \tag{1}$$

Here, $\mathbf{w}_i$ is the rendering weight, σ_i and c_i denote the density and color at the sampled point i, respectively and δ_i is the distance between adjacent samples along the ray. The density value is approximated from the SDF s, with learnable parameter α, β, as follows:

$$\sigma(s) = \begin{cases} \frac{1}{2} \exp\left(\frac{s}{\beta}\right) \cdot \alpha & \text{if } s \leq 0 \\ \left(1 - \frac{1}{2} \exp\left(-\frac{s}{\beta}\right)\right) \cdot \alpha & \text{if } s > 0 \end{cases} \tag{2}$$

3.2. *S-VolSDF*

Implicit neural 3D representations usually require dense images, since their per-scene optimization can be seen as a trial-and-error process to determine the underlying 3D structures. Therefore, given sparse training views as supervision, neural rendering models often fit the training views flawlessly while the underlying geometry can be vastly incorrect [42]. This can be understood as a local optimum and is known as the shape-radiance ambiguity problem [82]. In Fig. 1, we demonstrate it experimentally by training VolSDF [74] using 3 views only. As shown in Fig. 1, VolSDF completely fails to estimate geometry.

We propose to combine information from MVS to make neural rendering models correspondence-aware. The challenge lies in effectively incorporating noisy MVS predictions into VolSDF [74]. We propose two steps:

Soft Consistency. Instead of the hard consistency constraints imposed by estimating the depth of each point, we impose soft consistency constraints by operating directly on the probability volumes: In MVS, depth maps are typically obtained by applying *argmax* on the probability volume along each view direction. Then, photometric and geometric consistency checks [71] are used to filter out depth outliers before fusing the depth maps into a point cloud. *argmax* works well when dense inputs are available, but in the case of sparse inputs, the correct depth is often not assigned the highest probability. As a result, incorrect depths

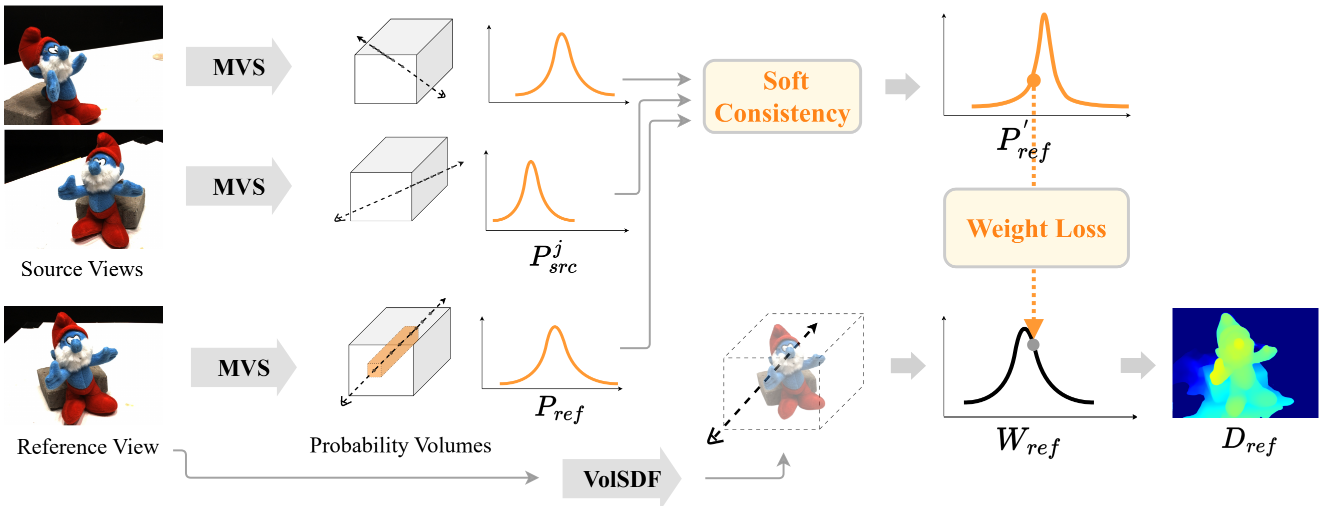


Figure 3. **Overview**. We propose to use probability volumes, obtained from multi-view stereo (MVS) models, to supervise the rendering weight estimated by VolSDF [74]. We apply a soft consistency check to refine the volumes. The weight loss function ensures consistency between the probability volume and the rendering weight. This process allows us to use the reconstructed depth information provided by VolSDF to guide the depth hypothesis sampling in the MVS models, as depicted in Fig. 2.

introduced by *argmax* will be filtered out by consistency checks, resulting in an incomplete reconstruction.

Alternatively, we propose directly computing consistency measures on the probability volumes. The *reference* view is the image, the depth of which we want to determine. The other images are the *source* views. By applying MVS to these views, we obtain probability volumes. Then, we multiply each probability value $\mathbf{P}_{ref}(\mathbf{x})$ in the reference probability volume with 3D position $\mathbf{x}$, with the sum of $\mathbf{P}^j_{src}(\mathbf{x})$ at the same location, to compute a new consistency weighted probability volume. $\mathbf{P}^j_{src}(\mathbf{x})$ is interpolated from the probability volumes of the source views. We demonstrate that this multiplication works adequately in our ablation study in Sec. 4.4. However, significant errors in depth still appear in challenging sparse-input scenarios.

Noise-Tolerant Loss. We further propose a noise-tolerant weight loss that utilizes the noisy probability volume to improve the reconstruction of VolSDF [74]. Given points sampled along a viewing ray, we notice that $\mathbf{P}$ and $\mathbf{w}$ in Eq. (1) actually have the same meaning: their normalized values along the ray/depth both form a depth probability mass function, which can also be seen as *the probability that a correspondence exists*. The larger the value of $\mathbf{w}$, the more likely it is that this point is visible in multiple views (i.e. a correspondence between pixels in different images). Instead of directly checking for consistency between the different probability volumes, we use them (in the form of $\mathbf{w}$) during neural rendering optimization. Thus, we leverage the smoothness of neural implicit models and combine the global consistency guaranteed by volumetric rendering.

Specifically, we use our consistency weighted probability volume as supervision to regularize $\mathbf{w}$ in the volume rendering Eq. (1). Based on [46], we can think of all possible 3D points on a ray as interior or exterior to the object (i.e. a binary classification problem). Thus, the MVS probability volume becomes a set of noisy positive labels for the rendering weights (i.e. occupancy values) with confidence from soft consistency. Hence, we have a classification problem that allows the use of cross entropy loss to optimize neural rendering methods. However, as shown in [85, 54], the cross entropy loss is sensitive to noisy labels. Based on insights from [85], we adopt a generalized cross entropy loss in Eq. (3) to reduce the penalty on false positive MVS predictions and thus increase optimization tolerance to noise. The noise tolerance level can be controlled by parameter q, where the generalized cross entropy loss is equivalent to the cross entropy loss when q approaches 0 [85], and to the Mean Absolute Error (MAE) loss when $q = 1$. Our noise-tolerant weight loss is shown in Eq. (3).

$$\begin{aligned} \mathcal{L}_{weight} &= \sum_{\mathbf{x}\in\mathcal{X}} \frac{1-\mathbf{w}(\mathbf{x})^q}{q} \cdot \mathbf{P}'_{ref}(\mathbf{x}), \\ \text{where} \quad \mathbf{P}'_{ref}(\mathbf{x}) &= \mathbf{P}_{ref}(\mathbf{x}) \cdot \sum_j \mathbf{P}^j_{src}(\mathbf{x}), \end{aligned} \tag{3}$$

$\mathbf{w}(\mathbf{x})$ is the rendering weight predicted by the neural rendering model at the sampled location $\mathbf{x}$ along a ray in the reference view, $\mathbf{P}_{ref}$ is the probability volume in the reference view, and $\mathbf{P}^j_{src}$ is the probability volume of a source view. In this way, we are essentially optimizing correspondences across images in a globally consistent and noise-tolerant way. In spirit, this is similar to finding a graph-cut in a volume of correspondence costs described in [58].

Coarse-to-fine MVS Reconstruction. As shown in Fig. 2, we incorporate our method into three coarse to fine MVS models [16, 9, 21]. We use the first coarse stage MVS probability volume to guide VolSDF [74] optimization. After that, we use the depth map obtained from VolSDF and replace the original depth map estimated by the coarse stage MVS model to remove the noise in the coarse stage MVS depth. We then follow the same protocol as in coarse-to-fine MVS models: use the depth map estimated from the coarse stage to guide the sampling range of the depth candidates of the next, finer stage in MVS. Because our coarse guidance depth map is more complete and accurate, the next stage MVS depth estimation is simpler. Therefore, we only use half of the depth search width in the finer stages compared to the default search width used in MVS models. Our surface reconstruction is more complete and still accurate compared to MVS models. Furthermore, as we show in Tab. 1, our method can be effortlessly incorporated into most coarse-to-fine MVS models and achieves considerable improvement compared to standalone MVS models.

Optimization. We use the same loss functions as VolSDF [74], along with our weight loss and sparsity regularization:

$$\mathcal{L}_{\text{sparse}} = \frac{1}{\|\mathbb{Q}\|} \sum_{r \in \mathbb{Q}} 1/(d_r + \epsilon), \tag{4}$$

where d_r are predicted depths and $\mathbb{Q}$ are rays without MVS supervision ($\sum \mathbf{P}'(\mathbf{x}) \approx 0$). We encourage sparsity by maximizing depth values. $\mathcal{L}_{\text{sparse}}$ is only used in the first 200 steps, along with heavily Gaussian-smoothed images as photometric supervision to suppress floating surfaces.

Rendering. For coordinate-based MLPs, fitting high-frequency details and maintaining good geometry simultaneously is challenging [74]. Since our method produces reasonably good geometry, we experiment with a simple image-based rendering approach [13, 8, 3] in testing to warp nearby view pixels based on predicted depth maps to synthesize novel views. In areas where there are no valid pixels to warp (i.e., the geometric consistency check between the rendered depths of the novel view and input views fails), we use rendered colors. A 4-level Laplacian pyramid [4] is used to smoothly blend the warped pixels. Our method with image-based rendering is denoted as **Ours**$_{IR}$.

4. Experiments

We evaluate our method on complex multi-view 3D surface reconstruction tasks, using two datasets: DTU [1] and BlendedMVS [73], both featuring real objects with diverse materials captured from multiple views. We demonstrate the superiority of our approach over prior work through quantitative and qualitative evaluation (Sec. 4.3). Furthermore, we conduct extensive ablation studies to verify the effectiveness of our design choices (Sec. 4.4).

4.1. Experimental Settings

Datasets. For the DTU dataset [1], we combine the scans used in [75, 74, 76] with the ones used in conventional MVS settings [16, 71], and remove the training scans of common MVS models. Our primary experiments are on three-view 3D reconstruction. Similar to PixelNeRF [76], we use views 25, 22, and 28 for three-view reconstruction. We further test on 6 and 9 input views with consistent improvements in performance[1]. For the BlendedMVS dataset [73], we select 9 challenge scenes, following [74]. For each scene, we select a set of sparse input views (i.e. 3 images) with a relatively wide baseline, similar to the setting in the DTU dataset. The image resolution is set to 768 × 576 for both the DTU and BlendedMVS datasets. We use foreground masks from [42, 75] following [42, 37] for evaluation.

Metrics. For surface reconstruction, we follow the standard evaluation protocol in [1, 75, 74] and report the Chamfer distance (in mm) of the output point clouds with ground truth point clouds. For novel view synthesis, we adopt the mean of PSNR, structural similarity index (SSIM) [64], and the LPIPS perceptual metric [83].

Implementation details. We experiment mainly using CasMVSNet [21] to obtain the cascade probability volume. We notice that, given only 3 input views, the default plane sweep settings (48, 32, and 8 depth hypotheses with interval widths 4, 2, and 1 respectively) do not retain fine details very well. We change them to 192, 32, and 8 depth hypotheses with intervals 1, 0.5, and 0.5 respectively. We are able to make the finer stage depth search interval widths much smaller because our method produces more complete and accurate coarse depth maps. The batch size is 512 rays. The q in our weight loss Eq. (3) is 0.5 in all experiments. We optimize each scene for 100K steps. Before fusing the depth maps output by the MVS model into a point cloud, standard photometric and geometric consistency [71] checks based on probability values and depth errors are adopted.

4.2. Baselines

Neural Rendering Methods. We compare against state-of-the-art generic neural rendering methods, including IBRNet [62], MVSNeRF [6], GeoNeRF [25], and SparseNeuS [37]. We fine-tune IBRNet and SparseNeuS using three input images for each scene for 20K and 10k iterations, respectively. We only report non-fine-tuned results for MVSNeRF and GeoNeRF because our attempts to fine-tune using 3 images did not succeed due to the inherent difficulty of the task, consistent with [42, 37]. Additionally, we compare our method with per-scene optimization based neural surface reconstruction methods, NeuS [61] and VolSDF [74],

[1]Results for the 6 and 9 image scenarios are in supplementary.

Scan	21	24	34	37	38	40	82	106	110	114	118	**Mean**
IBRNet$_{ft}$ [62]	3.40	3.54	3.13	6.78	3.32	4.80	3.48	2.59	3.93	1.23	2.74	3.54
MVSNeRF [6]	2.07	2.35	1.23	3.87	1.36	2.40	2.23	1.64	1.76	0.65	1.86	1.95
GeoNeRF [25]	2.13	3.04	1.00	3.93	1.20	2.46	2.32	1.82	2.21	0.79	1.67	2.06
SparseNeuS$_{ft}$ [37]	5.26	4.93	5.59	7.04	5.18	7.38	4.78	4.58	5.61	4.55	5.61	5.51
NeuS [61]	4.52	3.33	3.03	4.77	1.87	4.35	1.89	4.18	5.46	1.09	2.40	3.36
VolSDF [74]	4.54	2.61	1.51	4.05	1.27	3.58	3.48	2.62	2.79	0.52	1.10	2.56
TransMVSNet [16]	3.39	3.61	1.55	4.24	1.95	3.06	3.45	2.94	3.81	1.67	2.34	2.92
[16] + **Ours**	1.91	2.08	0.98	2.94	1.21	1.90	2.70	1.56	1.99	1.09	1.35	1.80
UCSNet [9]	2.57	3.01	1.82	4.07	1.62	3.10	2.49	1.93	1.27	0.68	1.59	2.20
[9] + **Ours**	**1.89**	2.12	1.24	3.17	1.07	2.07	1.38	**1.24**	0.78	0.54	1.16	1.52
CasMVSNet[21]	2.40	3.07	1.23	3.27	1.35	2.76	1.82	1.72	1.30	0.70	1.44	1.92
[21] + **Ours**	1.96	**1.99**	**0.74**	**2.58**	**0.95**	**1.47**	**1.37**	1.32	**0.54**	**0.51**	**1.03**	**1.32**

Table 1. Quantitative results on 3D reconstruction for the DTU dataset. "+ Ours" means that we use the cited MVS algorithm as the probability volume builder and optimize using our method. The metric is the Chamfer distance (lower is better).

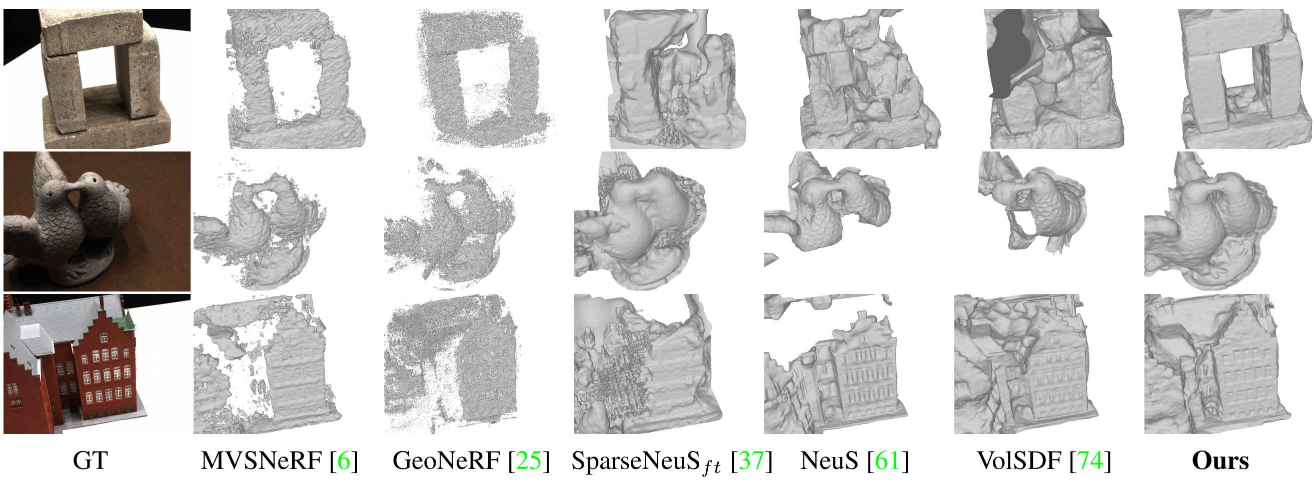


Figure 4. 3D reconstruction results of neural rendering methods on DTU. Our results appear more complete and accurate.

with VolSDF being the neural rendering model used to refine MVS predictions in our method. For fair comparison with MVS, we only maintain the foreground depth maps generated by neural rendering techniques by applying standard geometric consistency checks and ground truth masks. We merge the depth maps into a point cloud for evaluation [71].

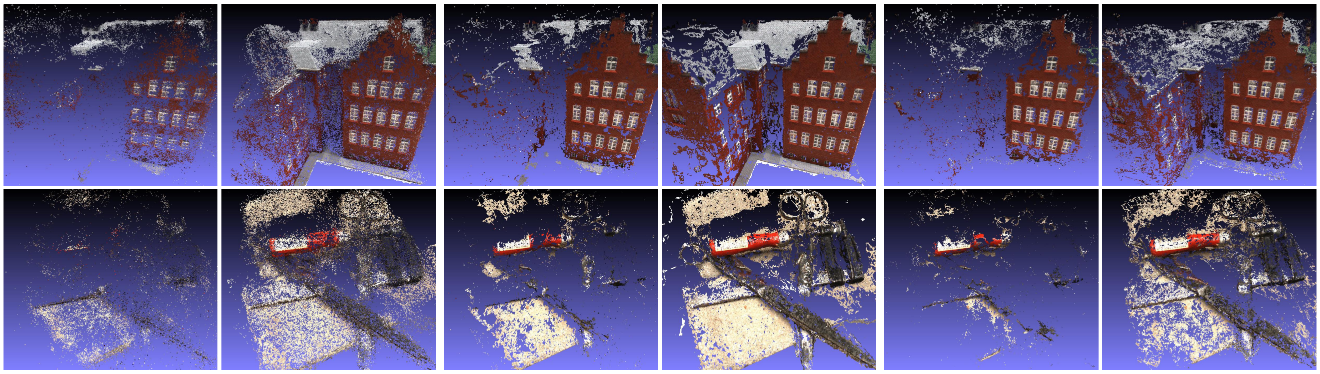
MVS Methods. To evaluate the generalizability of our method, we incorporate it into three state-of-the-art coarse to fine MVS models: CasMVSNet [21], UCSNet [9], and TransMVSNet [16]. All MVS networks are pre-trained only on DTU [1] with ground-truth depth as supervision and are frozen during per-scene optimization.

4.3. Comparisons

3D Reconstruction. Our approach surpasses state-of-the-art techniques in 3D reconstruction, as demonstrated by its superior performance on both the DTU [1] and BlendedMVS datasets [73] (Tab. 1 and Tab. 2). We show meshes extracted from neural rendering method outputs in Fig. 4 and Fig. 7.

As shown in Fig. 4, VolSDF [74] and NeuS [61] show suboptimal performance due to the weak photometric constraint in resolving the shape-radiance ambiguity. Fine-tuning SparseNeuS [37] can lead to degenerate results, especially on the BlendedMVS dataset, so we only report its performance on DTU. Fine-tuned IBRNet [62] performs worse than methods using stronger MVS priors such as MVSNeRF [6] and GeoNeRF [25]. Although MVSNeRF and GeoNeRF demonstrate impressive performance, they still fall short compared to our method (see Fig. 8).

As shown in Tab. 1 and Fig. 5, MVS models coupled with our noise-tolerant optimization perform much better than MVS models or VolSDF [74] alone. Thus, our method



TransMVSNet [16] [16]+Ours CasMVSNet [21] [21]+Ours UCSNet [9] [9]+Ours

Figure 5. Point cloud visualization on DTU. Results improve in all combinations of our method with different MVS models.

GT IBRNet$_{ft}$ [62] MVSNeRF [6] GeoNeRF [25] NeuS [61] VolSDF [74] **Ours** **Ours$_{IR}$**

Figure 6. Our method appears to be more accurate in novel view synthesis on DTU.

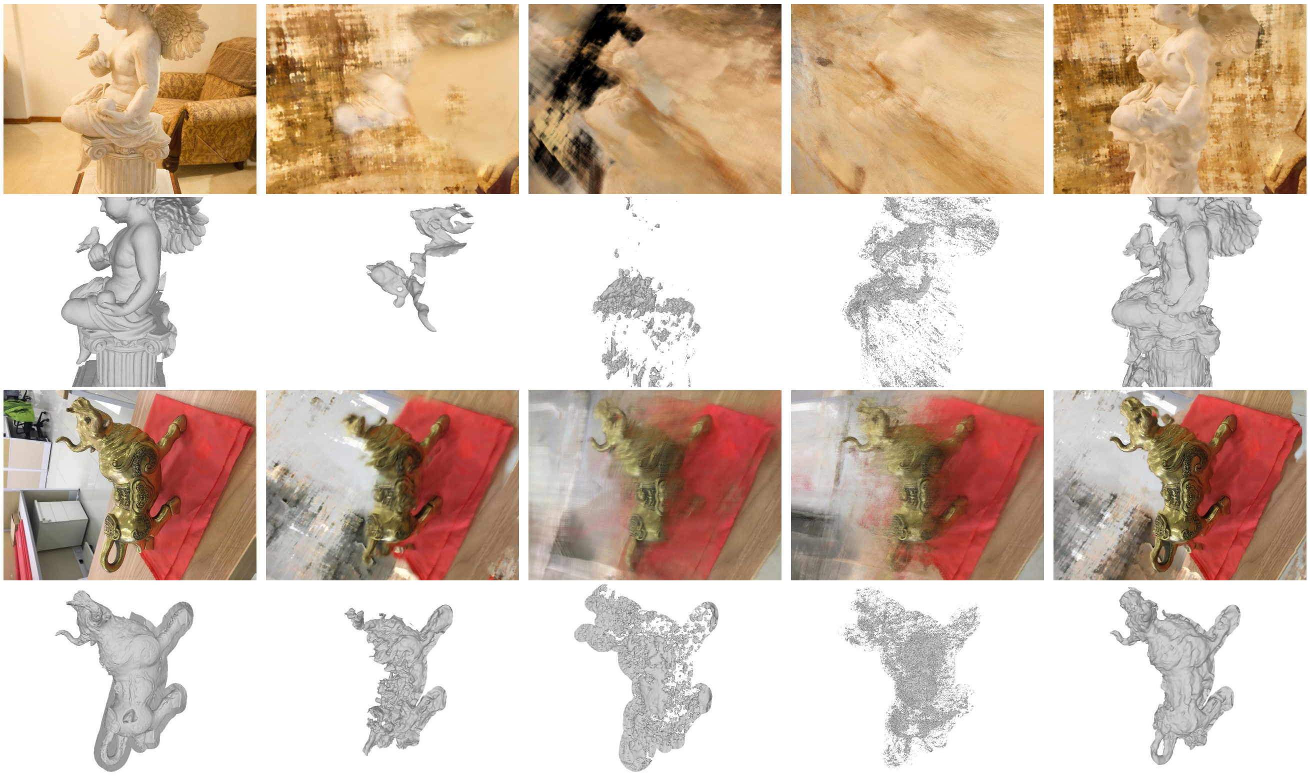
Scene	Doll	Egg	Head	Angel	Bull	Robot	Dog	Bread	Camera	**Mean**
MVSNeRF [6]	5.3	-16.8	-17.7	38.2	13.8	11.9	-0.3	12.3	8.0	6.1
GeoNeRF [25]	29.3	21.4	11.5	37.6	-0.6	-15.1	13.7	21.4	11.5	14.5
CasMVSNet [21]	32.9	47.1	17.3	45.9	11.3	11.5	33.3	19.2	30.1	27.6
Ours	**35.0**	**58.8**	**38.5**	**54.7**	**33.4**	**23.9**	**33.7**	**64.4**	**43.4**	**42.9**

Table 2. BlendedMVS 3D reconstruction results. Since there are no units in BlendedMVS, we report relative improvement (in %) over VolSDF [74] in terms of Chamfer distance.

can be treated as a general module that can be plugged into other MVS methods and improve their performance.

With the introduction of MVS information, we enable fast per-scene surface optimization. Our output surface reconstruction after 10-15 minutes of training (on an NVIDIA A5000 GPU) is already better than the reconstruction of the fully trained VolSDF [74] (typically 4-10 hours). More specifically, on DTU, we obtain 39% better Chamfer distance over the fully-trained VolSDF after 15 minutes of optimization, with our final model achieving a 48% improvement. Please refer to the supplementary for more details.

Novel View Synthesis. Our method excels at improving geometry, yet also demonstrates competitive performance in novel view synthesis (as shown in Tab. 3 and Tab. 4). Fig. 6 and Fig. 7 illustrate improved view synthesis results compared to other methods, suggesting our method's capacity to better disentangle geometry and texture. Also, adding the image interpolation to the rendering process greatly enhances LPIPS, while slightly improving PSNR and SSIM, by incorporating more details, as demonstrated in Tab. 3, Tab. 4, and Fig. 6.



GT VolSDF [74] MVSNeRF [6] GeoNeRF [25] **Ours**$_{IR}$

Figure 7. 3D reconstruction and novel view synthesis comparisons on BlendedMVS. Our results appear more complete and accurate.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
IBRNet$_{ft}$ [62]	15.71	0.759	0.295
MVSNeRF [6]	18.37	0.818	0.254
GeoNeRF [25]	19.45	0.837	0.220
NeuS [61]	15.34	0.753	0.313
VolSDF [74]	16.99	0.786	0.332
Ours	20.21	0.820	0.321
Ours$_{IR}$	**20.58**	**0.855**	**0.157**

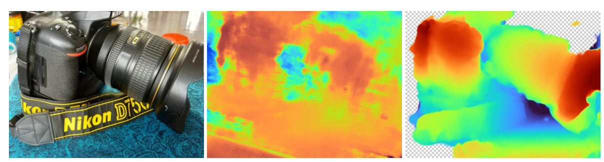
Table 3. Novel view synthesis comparisons on DTU.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
MVSNeRF [6]	14.99	0.866	0.164
GeoNeRF [25]	17.09	0.886	0.139
VolSDF [74]	14.47	0.860	0.182
Ours	16.97	0.893	0.154
Ours$_{IR}$	**17.26**	**0.906**	**0.105**

Table 4. Novel view synthesis comparisons for BlendedMVS.

4.4. Ablation Study

We conduct ablation studies on the DTU dataset (Tab. 5). First, we show that using only the soft consistency constraints without additional optimization still improves the



GT MVSNeRF [6] Ours

Figure 8. Depth map predictions on BlendedMVS using MVSNeRF [6] and our method. Improved depths are an illustration of better geometry-appearance disentanglement.

reconstruction result. This supports our assumption that the probability volumes contain more information than lossy depth maps obtained from an *argmax* operation. Second, to evaluate the effectiveness of our weight loss, we replace our loss with the mean squared error (MSE) between the reconstructed depth from VolSDF and the geo-consistency filtered depth map obtained from MVS, similar to DS-NeRF [14]. Third, replacing the probability volumes with the depth maps as input, led to worse performance[2]. Finally, we replace our weight loss with cross entropy loss, showing that generalized cross entropy loss is indeed noise-tolerant. Due to the trade-off between accuracy and completeness in

[2]We set the probability to be 1 only at the depth prediction location.

point cloud filtering, we use Chamfer distance as the metric, following [75, 74]. See supplementary for more details.

Method	Chamfer ↓
VolSDF [74]	2.558
CasMVSNet [21]	1.920
Ours	**1.320**
only soft consistency	1.711
MSE loss [14]	1.792
w/o probability volume	1.543
w/o GCE loss	1.534

Table 5. Ablation studies for the DTU dataset. All rows except the first three are our model with different ablated components.

5. Conclusions

We presented *S-VolSDF*, a novel approach to recover underlying geometry from sparse input views. Neural rendering optimization mainly relies on dense input images so that it can use trial-and-error mechanisms for reconstruction. Hence, its performance drops considerably with sparse inputs. We regularized the weight distribution with a refined probability volume obtained from MVS algorithms. We further made our method noise-tolerant by applying a generalized cross entropy loss. Our experiments show that our model not only outperforms neural rendering models but also significantly boosts the performance of MVS algorithms.

Discussion and Limitations. While our method is capable of refining the probability volumes of the finer stages of MVS, we notice diminishing improvement because there is not much noise left in these stages. We include an ablation study on this in the supplementary material. While neural rendering models are able to deal with non-opaque, textureless, or glossy surfaces, our introduction of MVS reduces this ability. This is an interesting area of research, particularly in the context of few-view reconstruction.

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120(2):153–168, 2016. 2, 5, 6

[2] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 2

[3] Chris Buehler, Michael Bosse, Leonard McMillan, Steven Gortler, and Michael Cohen. Unstructured lumigraph rendering. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 425–432, 2001. 5

[4] Peter J Burt and Edward H Adelson. The laplacian pyramid as a compact image code. In *Readings in computer vision*, pages 671–679. Elsevier, 1987. 5

[5] Neill DF Campbell, George Vogiatzis, Carlos Hernández, and Roberto Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *European Conference on Computer Vision*, pages 766–779. Springer, 2008. 2

[6] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021. 3, 5, 6, 7, 8

[7] Rui Chen, Songfang Han, Jing Xu, and Hao Su. Point-based multi-view stereo network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1538–1547, 2019. 1, 2

[8] Shenchang Eric Chen and Lance Williams. View interpolation for image synthesis. In *Proceedings of the 20th annual conference on Computer graphics and interactive techniques*, pages 279–288, 1993. 5

[9] Shuo Cheng, Zexiang Xu, Shilin Zhu, Zhuwen Li, Li Erran Li, Ravi Ramamoorthi, and Hao Su. Deep stereo using adaptive thin volume representation with uncertainty awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2524–2534, 2020. 1, 2, 5, 6, 7

[10] Julian Chibane, Aayush Bansal, Verica Lazova, and Gerard Pons-Moll. Stereo radiance fields (srf): Learning view synthesis for sparse views of novel scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7911–7920, 2021. 3

[11] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 303–312, 1996. 2

[12] 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. 2

[13] Paul E Debevec, Camillo J Taylor, and Jitendra Malik. Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 11–20, 1996. 5

[14] 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. 2, 8, 9

[15] Arnab Dey, Yassine Ahmine, and Andrew I Comport. Mip-nerf rgb-d: Depth assisted fast neural radiance fields. *arXiv preprint arXiv:2205.09351*, 2022. 2

[16] Yikang Ding, Wentao Yuan, Qingtian Zhu, Haotian Zhang, Xiangyue Liu, Yuanjiang Wang, and Xiao Liu. Transmvsnet: Global context-aware multi-view stereo network with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8585–8594, 2022. 1, 2, 5, 6, 7

[17] Yasutaka Furukawa, Brian Curless, Steven M Seitz, and Richard Szeliski. Reconstructing building interiors from images. In *2009 IEEE 12th international conference on computer vision*, pages 80–87. IEEE, 2009. 2

[18] Yasutaka Furukawa, Carlos Hernández, et al. Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015. 2

[19] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2009. 2

[20] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 873–881, 2015. 2

[21] 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. 1, 2, 5, 6, 7, 9

[22] Carlos Hernández, George Vogiatzis, and Roberto Cipolla. Probabilistic visibility for multi-view stereo. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 2

[23] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5885–5894, 2021. 2

[24] Yue Jiang, Dantong Ji, Zhizhong Han, and Matthias Zwicker. Sdfdiff: Differentiable rendering of signed distance fields for 3d shape optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1251–1261, 2020. 2

[25] Mohammad Mahdi Johari, Yann Lepoittevin, and François Fleuret. Geonerf: Generalizing nerf with geometry priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18365–18375, 2022. 3, 5, 6, 7, 8

[26] James T Kajiya and Brian P Von Herzen. Ray tracing volume densities. *ACM SIGGRAPH computer graphics*, 18(3):165–174, 1984. 2, 3

[27] Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine. *Advances in neural information processing systems*, 30, 2017. 2

[28] Petr Kellnhofer, Lars C Jebe, Andrew Jones, Ryan Spicer, Kari Pulli, and Gordon Wetzstein. Neural lumigraph rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4287–4297, 2021. 2

[29] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE international conference on computer vision*, pages 66–75, 2017. 2

[30] Johannes Kirmayr and Philipp Wulff. Dynamic nerf on rgb-d data. Jul 2022. 2

[31] Kiriakos N Kutulakos and Steven M Seitz. A theory of shape by space carving. *International journal of computer vision*, 38(3):199–218, 2000. 2

[32] Maxime Lhuillier and Long Quan. A quasi-dense approach to surface reconstruction from uncalibrated images. *IEEE transactions on pattern analysis and machine intelligence*, 27(3):418–433, 2005. 2

[33] Shaohui Liu, Yinda Zhang, Songyou Peng, Boxin Shi, Marc Pollefeys, and Zhaopeng Cui. Dist: Rendering deep implicit signed distance function with differentiable sphere tracing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2019–2028, 2020. 2

[34] Yuan Liu, Sida Peng, Lingjie Liu, Qianqian Wang, Peng Wang, Christian Theobalt, Xiaowei Zhou, and Wenping Wang. Neural rays for occlusion-aware image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7824–7833, 2022. 3

[35] Alex Locher, Michal Perdoch, and Luc Van Gool. Progressive prioritized multi-view stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3244–3252, 2016. 2

[36] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751*, 2019. 2

[37] Xiaoxiao Long, Cheng Lin, Peng Wang, Taku Komura, and Wenping Wang. Sparseneus: Fast generalizable neural surface reconstruction from sparse views. *arXiv preprint arXiv:2206.05737*, 2022. 3, 5, 6

[38] Keyang Luo, Tao Guan, Lili Ju, Haipeng Huang, and Yawei Luo. P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10452–10461, 2019. 1, 2

[39] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7210–7219, 2021. 2

[40] Nelson Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1(2):99–108, 1995. 2, 3

[41] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020. 2, 3

[42] 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. 2, 3, 5

[43] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3504–3515, 2020. 2

[44] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5589–5599, 2021. 2

[45] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 2

[46] Rui Peng, Rongjie Wang, Zhenyu Wang, Yawen Lai, and Ronggang Wang. Rethinking depth estimation for multi-view stereo: A unified representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8645–8654, 2022. 1, 4

[47] Barbara Roessle, Jonathan T Barron, Ben Mildenhall, Pratul P Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12892–12901, 2022. 2

[48] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2304–2314, 2019. 2

[49] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 2

[50] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 501–518. Springer, 2016. 2

[51] Steven M Seitz and Charles R Dyer. Photorealistic scene reconstruction by voxel coloring. *International Journal of Computer Vision*, 35(2):151–173, 1999. 2

[52] Sudipta N Sinha, Philippos Mordohai, and Marc Pollefeys. Multi-view stereo via graph cuts on the dual of an adaptive tetrahedral mesh. In *2007 IEEE 11th international conference on computer vision*, pages 1–8. IEEE, 2007. 2

[53] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[54] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 2, 4

[55] Mohammed Suhail, Carlos Esteves, Leonid Sigal, and Ameesh Makadia. Generalizable patch-based neural rendering. *arXiv preprint arXiv:2207.10662*, 2022. 3

[56] Engin Tola, Christoph Strecha, and Pascal Fua. Efficient large-scale multi-view stereo for ultra high-resolution image sets. *Machine Vision and Applications*, 23(5):903–920, 2012. 2

[57] Alex Trevithick and Bo Yang. Grf: Learning a general radiance field for 3d representation and rendering. *arXiv preprint arXiv:2010.04595*, 2020. 2

[58] George Vogiatzis, Philip HS Torr, and Roberto Cipolla. Multi-view stereo via volumetric graph-cuts. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 391–398. IEEE, 2005. 2, 5

[59] Guangcong Wang, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. *arXiv preprint arXiv:2303.16196*, 2023. 2

[60] Jiepeng Wang, Peng Wang, Xiaoxiao Long, Christian Theobalt, Taku Komura, Lingjie Liu, and Wenping Wang. Neuris: Neural reconstruction of indoor scenes using normal priors. *arXiv preprint arXiv:2206.13597*, 2022. 2

[61] 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. *arXiv preprint arXiv:2106.10689*, 2021. 1, 2, 5, 6, 7, 8

[62] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021. 3, 5, 6, 7, 8

[63] Xiaofeng Wang, Zheng Zhu, Fangbo Qin, Yun Ye, Guan Huang, Xu Chi, Yijia He, and Xingang Wang. Mvster: Epipolar transformer for efficient multi-view stereo. *arXiv preprint arXiv:2204.07346*, 2022. 1, 2

[64] 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

[65] Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, and Jie Zhou. Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5610–5619, 2021. 2

[66] Zizhuang Wei, Qingtian Zhu, Chen Min, Yisong Chen, and Guoping Wang. Aa-rmvsnet: Adaptive aggregation recurrent multi-view stereo network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6187–6196, 2021. 1, 2

[67] Qingshan Xu and Wenbing Tao. Multi-scale geometric consistency guided multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5483–5492, 2019. 2

[68] Qingshan Xu and Wenbing Tao. Learning inverse depth regression for multi-view stereo with correlation cost volume. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12508–12515, 2020. 1, 2

[69] Jianfeng Yan, Zizhuang Wei, Hongwei Yi, Mingyu Ding, Runze Zhang, Yisong Chen, Guoping Wang, and Yu-Wing Tai. Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. In *European conference on computer vision*, pages 674–689. Springer, 2020. 1, 2

[70] Jiayu Yang, Wei Mao, Jose M Alvarez, and Miaomiao Liu. Cost volume pyramid based depth inference for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4877–4886, 2020. 1, 2

[71] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 767–783, 2018. 1, 2, 3, 5, 6

[72] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5525–5534, 2019. 1, 2

[73] 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

[74] 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, 3, 4, 5, 6, 7, 8, 9

[75] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33:2492–2502, 2020. 1, 2, 5, 9

[76] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 3, 5

[77] Zehao Yu and Shenghua Gao. Fast-mvsnet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1949–1958, 2020. 1, 2

[78] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *arXiv preprint arXiv:2206.00665*, 2022. 2

[79] Christopher Zach, Thomas Pock, and Horst Bischof. A globally optimal algorithm for robust tv-l 1 range image integration. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007. 2

[80] Jingyang Zhang, Yao Yao, Shiwei Li, Zixin Luo, and Tian Fang. Visibility-aware multi-view stereo network. *arXiv preprint arXiv:2008.07928*, 2020. 1, 2

[81] 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. 2

[82] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 1, 3

[83] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[84] Xudong Zhang, Yutao Hu, Haochen Wang, Xianbin Cao, and Baochang Zhang. Long-range attention network for multi-view stereo. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3782–3791, 2021. 1, 2

[85] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018. 2, 4

[86] Jie Zhu, Bo Peng, Wanqing Li, Haifeng Shen, Zhe Zhang, and Jianjun Lei. Multi-view stereo with transformer. *arXiv preprint arXiv:2112.00336*, 2021. 1, 2

[87] Yiming Zuo and Jia Deng. View synthesis with sculpted neural points. *arXiv preprint arXiv:2205.05869*, 2022. 2