Learning Unsigned Distance Functions from Multi-view Images with Volume Rendering Priors

Anonymous ECCV 2024 Submission

Paper ID #6611

Abstract. Unsigned distance functions (UDFs) have been a vital representation for open surfaces. With different differentiable renderers, current methods are able to train neural networks to infer a UDF by minimizing the rendering errors on the UDF to the multi-view ground truth. However, these differentiable renderers are mainly handcrafted, which makes them either biased on ray-surface intersections, or sensitive to unsigned distance outliers, or not scalable to large scale scenes. To resolve these issues, we present a novel differentiable renderer to infer UDFs more accurately. Instead of using handcrafted equations, our differentiable renderer is a neural network which is pre-trained in a data-driven manner. It learns how to render unsigned distances into depth images, leading to a prior knowledge, dubbed volume rendering priors. To infer a UDF for an unseen scene from multiple RGB images, we generalize the learned volume rendering priors to map inferred unsigned distances in alpha blending for RGB image rendering. Our results show that the learned volume rendering priors are unbiased, robust, scalable, 3D aware, and more importantly, easy to learn. We evaluate our method on both widely used benchmarks and real scenes, and report superior performance over the state-of-the-art methods.

Keywords: Unsigned distance function · Volume rendering · Implicit reconstruction

1 Introduction

Neural implicit representations have become a dominated representation in 3D computer vision. Using coordinate based deep neural networks, a mapping from locations to attributes at these locations like geometry [32,36], color [10,31], and motion [14] can be learned as an implicit representation. Signed distance function (SDF) [32] and unsigned distance function (UDF) [9] are widely used implicit representations to represent either closed surfaces [20,36] or open surfaces [16,51]. We can learn SDFs or UDFs from supervisions like ground truth signed or unsigned distances [4,32], 3D point clouds [8,28,40] or multi-view images [26,46]. Compared to SDFs, it is a more challenging task to estimate a UDF due to the sign ambiguity and the boundary effect, especially under a multi-view setting.

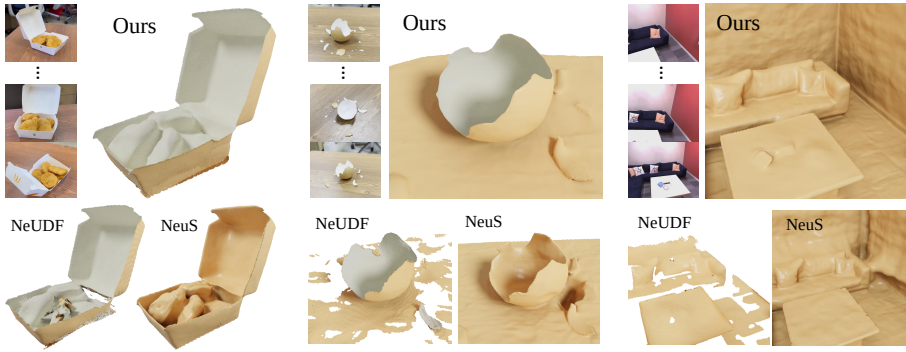


Fig. 1: We highlight our multi-view reconstruction results from UDFs learned on real-captured open surface scenes and indoor scenes. The two sides of a surface are colored in white and beige, respectively. Comparing with NeuS [41] and the state-of-the-art UDF reconstruction method NeUDF [25], our method does not produce artifacts and recovers more accurate and smooth geometries on both open and closed surfaces.

Recent methods [18, 25, 26, 29] mainly infer UDFs from multi-view images through volume rendering. Using different differentiable renderers, they can render a UDF into RGB or depth images which can be directly supervised by the ground truth images. These differentiable renderers are mainly hand-crafted equations which are either biased on ray-surface intersections, or sensitive to unsigned distance outliers, or not scalable to large scale scenes. These issues make them struggle with recovering accurate geometry. Fig. 2, 3 details these issues by comparing error maps on depth images rendered by different differentiable renderers. We render the ground truth UDF into depth images using different renderers from 100 different view angles, and report the average rendering error on each one of 55 shapes that are randomly sampled from each one of the 55 categories in ShapeNet [5] in Fig. 2. Using the latest differentiable renderers from NeuS-UDF [41] (using UDF as input to NeuS), NeUDF [25] and NeuralUDF [26], the rendered depth images and their error maps in Fig. 3 (a) to (c) show that these issues cause large errors even using

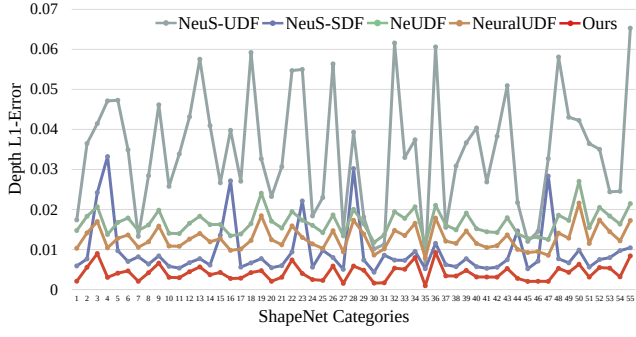


Fig. 2: Statistics of depth L1-error for various differentiable renderers. Each data point represents the mean depth L1-error computed between 100 predicted and GT depth maps of a random object from each category of ShapeNet.

the ground truth UDF as inputs. Therefore, how to design better differentiable renderers for UDF inference from multi-view images is still a challenge.

To resolve these issues, we introduce a novel differentiable renderer for UDF inference from multi-view images through volume rendering. Instead of handcrafted equations used by the latest methods [18, 25, 26, 29], we employ a neural network to learn to become a differentiable renderer in a data-driven manner. Using UDFs and depth images obtained from meshes as ground truth, we train the neural network to map a set of unsigned distances at consecutive locations along a ray into weights for alpha blending, so that we can render depth images, and produce the rendering errors to the ground truth as a loss. We make the neural network observe different variations of unsigned distance fields during training, and learn the knowledge of volume rendering with unsigned distances by minimizing the rendering loss. The knowledge we call *volume rendering prior* is highly generalizable to infer UDFs from multi-view RGB images in unobserved scenes. During testing, we use the pre-trained network as a differentiable renderer during alpha blending. It renders unsigned distances inferred by a UDF network into RGB images which can be supervised by the observed RGB images for UDF inference. Our results in Fig. 2 and Fig. 3 (e) show that we produce the smallest rendering errors among all differentiable renderers for UDFs, which is even more accurate than NeuS-SDF [41] (rendering with ground truth SDF) in Fig. 3 (d). Extensive experiments in our evaluations show that the learned volume rendering priors are unbiased, robust, scalable, 3D aware, and more importantly, easy to learn. We conduct evaluations in both widely used benchmarks and real scenes, and report superior performance over the state-of-the-art methods. Our contributions are listed below,

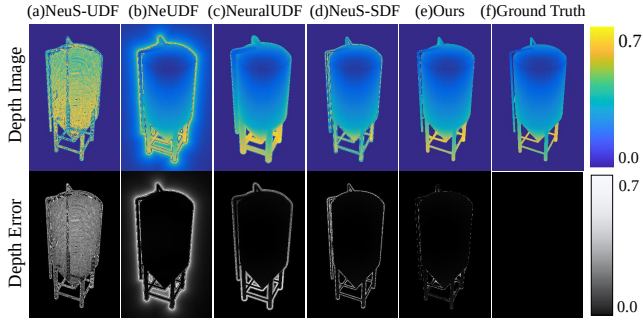


Fig. 3: Comparisons of estimated depth images and depth error maps among different differentiable renderers on one shape from the category of "tower" in ShapeNet.

- We introduce volume rendering priors to infer UDFs from multi-view images. Our prior can be learned in a data-driven manner, which provides a novel perspective to recover geometry with prior knowledge through volume rendering.
- We propose a novel deep neural network and learning scheme, and report extensive analysis to learn an unbiased differentiable renderer for UDFs with robustness, scalability, and 3D awareness.

– We report the state-of-the-art reconstruction accuracy from UDFs inferred from multi-view images on widely used benchmarks and real image sets.

2 Related Work

Multi-view 3D reconstruction. Multi-view 3D reconstruction aims to reconstruct 3D shapes from calibrated images captured from overlapping viewpoints. The key idea is to leverage the consistency of features across different views to infer the geometry. MVSNet [45] is the first to introduce the learning-based idea into traditional MVS methods. Following studies explore the potential of MVSNet in different aspects, such as training speed [43, 48], memory consumption [15, 44], network structure [7, 13] and generalization [50]. These techniques produce depth maps or 3D point clouds. To obtain meshes as final 3D representations, additional procedures such as TSDF-fusion [11] or classic surface reconstruction [21] methods are used, which is complex and not intuitive.

Learning SDFs from Multi-view Images. Instead of 3D point clouds estimated by MVS methods, recent methods [41, 42, 46] directly estimate SDFs through volume rendering from multi-view images for continuous surface representations. The widely used strategy is to render the estimated SDF into RGB images [12, 24, 33] or depth images [2, 39, 49] which can be supervised by the ground truth images. The key to make the whole procedure differentiable is various differentiable renderers [2, 30, 41] which transform signed distances into weights for alpha blending during rendering. Some methods modify the rendering equations to use more 2D supervisions like normal maps [38], detected planes [17], and segmentation maps [22] to pursue higher reconstruction efficiency. Similarly, some methods [37, 53, 54] jointly estimate camera poses and geometry in the context of SLAM. However, the SDFs that these methods aim to learn are only for closed surfaces, which is limited to represent open surfaces.

Learning UDFs from Multi-view Images. Different from SDFs, UDFs are able to represent open surfaces. Recent methods [18,25,26,29] design different differentiable renderes to learn UDFs through multi-view images. NeuralUDF [26] predicts the first intersection along a ray and flips the UDFs behind this point to use the differentiable renderer of NeuS [41]. NeUDF [25] proposes an inverse proportional function mapping UDF to rendering weights. NeAT [29] learns an additional validity probability net to predict the regions with open structures, while [18] proposes a bell-shaped weight function that maps UDF to density, inspired by HF-NeuS [42].

The differentiable renderers introduced by these methods mainly get formulated into handcrafted equations which are biased on ray-surface intersections, sensitive to unsigned distance outliers, and not 3D aware. We resolve this issue by introducing a learning-based differentiable renderer which learns and generalizes a volume rendering prior for robustness and scalability. The ideas of learnable neural rendering frameworks are also introduced in [1, 6, 23].

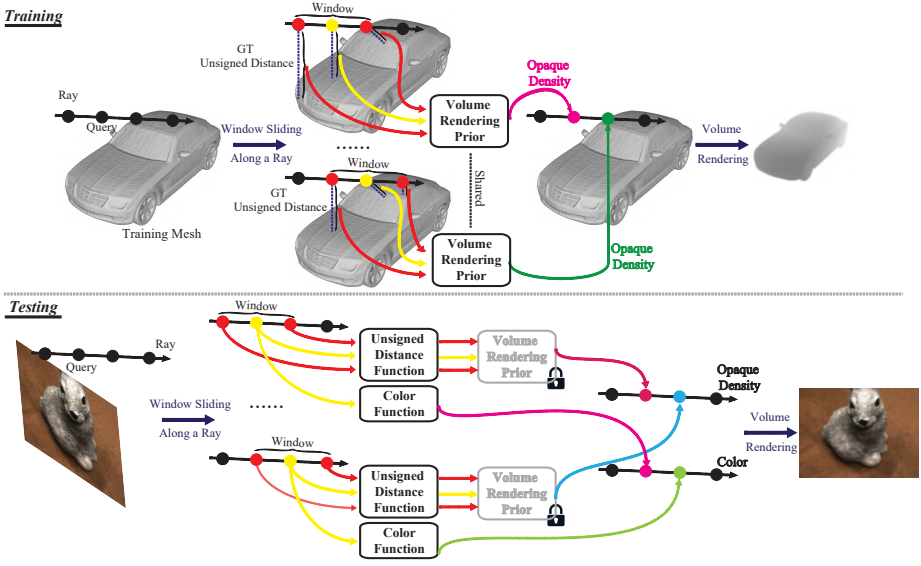


Fig. 4: Overview of our method. In the training phase, our volume rendering prior takes sliding windows of GT UDFs from training meshes as input, and outputs opaque densities for alpha blending. The parameters are optimized by the error between rendered depth and ground truth depth maps. During the testing phase, we freeze the volume rendering prior and use ground truth multi-view RGB images to optimize a randomly initialized UDF field.

3 Methods

Problem Statement. Given a set of J images $\{I_j\}_{j=1}^{J}$, we aim to infer a UDF f_u which predicts an unsigned distance u for an arbitrary 3D query q. We formulate the UDF as $u = f_u(q)$. With the learned f_u, we can extract the zero level set of f_u as a surface using algorithms similar to the marching cubes [16,51].

Overview. We employ a neural network to learn f_u by minimizing rendering errors to the ground truth. We shoot rays from each view I_j, sample queries q along each ray, and get unsigned distance prediction u from f_u to calculate weights w for alpha blending in volume rendering. At the same time, we train a color function c which predicts the color at these queries q as $c = f_c(q)$. The accumulation of c with weights w along the ray produces a color at the pixel.

Current differentiable renderers [18,25,26,29] transform u into w using handcrafted equations. Instead, we train a neural network to approximate this function f_w in a data-driven manner, as illustrated in Fig. 4. During training, we push f_w to produce ideal weights for rendering depth images that are as similar to the supervision $\{D_a^h\}_{a=1}^{A}$ from the h-th shape as possible using the ground truth UDF, and more importantly, get used to various variations of unsigned distances along a ray. During testing, we use this volume rendering prior with fixed parameters θ_w of f_w. We leverage f_w to estimate an f_u from multi-view

RGB images $\{I_j\}$ of an unseen scene by minimizing rendering errors of RGB color through volume rendering.

Volume Rendering for UDFs. We render a UDF function f_u with a color function f_c into either RGB I' or depth D' images to compare with the RGB supervision $\{I_j\}$ or depth supervision $\{D_j\}$. Note that we do not use depth supervision $\{D_j\}$ during the UDF inference, but we include a depth supervision here to make the UDF rendering with learned priors self-contained.

From each posed view I_j, we sample some pixels and shoot rays starting at each pixel. Taking a ray V_k from view I_j for example, V_k starts from the camera origin o and points to a direction r. We hierarchically sample N points along the ray V_k, where each point is sampled at $q_n = o + d_n * r$ and d_n corresponds to the depth value of q_n on the ray. We can transform unsigned distances $f_u(q_n)$ into weights w_n which is used for color or depth accumulation along the ray V_k in volume rendering,

$$\begin{aligned} \sigma_n &= f_w(\{q_m\}_{m=1}^{M}, \{f_u(q_m)\}_{m=1}^{M}), \\ w_n &= \sigma_n \times \prod_{n'=1}^{n-1}(1-\sigma_{n'}) \\ I(k)' &= \sum_{n'=1}^{N} w_{n'} \times f_c(q_{n'}), \\ D(k)' &= \sum_{n'=1}^{N} w_{n'} \times d_{n'}, \end{aligned} \tag{1}$$

where q_m is one of M nearest neighbors of q_n along a ray, and σ_n is the opaque density that can be interpreted as the differential probability of a ray terminating at an infinitesimal particle at the location q_n. The latest methods model the weighting function f_w in handcrafted ways with $M = 2$ neighbors around q_n on the same ray. For instance, NeUDF [25] introduced an inverse proportional function to calculate opaque density from two adjacent queries, while NeuralUDF [26] used the same two queries to model both occlusion probability and opaque densities.

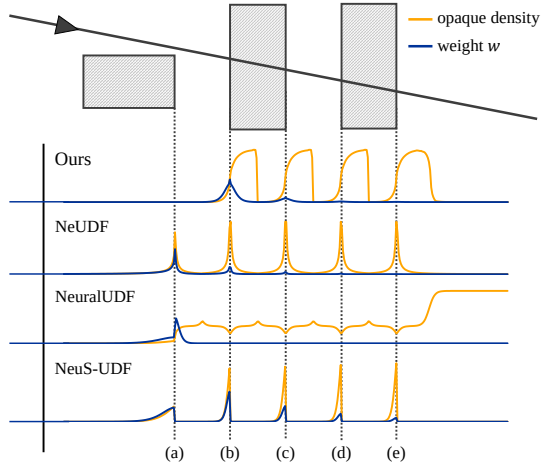


Fig. 5: Distribution of opaque densities and accumulated weights calculated by different baselines and predicted by our volume rendering priors. Our method is 3D aware and robust to unsigned distance changes at near-surface points while deriving unbiased volume rendering weights.

Although these differentiable renderers are unbiased at intersection of ray and surface and can render UDFs into images, they usually produce render errors on the boundaries on depth images, as shown by error

maps in Fig. 3 (b) and (c). These errors indicate that these handcrafted equations can not render correct depth even when using the ground truth unsigned distances as supervision.

Why do these handcrafted equations produce large rendering errors? Our analysis shows that being not 3D aware plays a key role in producing these errors. These handcrafted equations merely use M=2 neighboring points to perceive the 3D structure when calculating the opaque density at query q_n. Such a small window makes these equations merely have a pretty small receptive field, which makes them become sensitive to unsigned distance changes, such as the weight decrease at queries sampled on a ray that is passing by an object. Moreover, to maintain some characteristics like unbiasness and occlusion awareness, these equations are strictly handcrafted, which make them extremely hard to get extended to be more 3D aware by using more neighboring points as input. Another demerit comes from the fact that all rays need to use the same equation to model the opague, which is not generalizable enough to cover various unsigned distance combinations.

Fig. 5 illustrates issues of current methods. When a ray is approaching an object, handcrafted equations struggle to produce a zero opaque density at the location where the ray merely passes by an object but not intersects with it. This is also the reason why these methods produce large rendering errors on the boundary in Fig. 3. To resolve these issues, we introduce to train a neural network to learn the weight function f_w in a data-driven manner, which leads to a volume rendering prior. During training, the network observes huge amount of unsigned distance variations along rays, and learns how to map unsigned distances into weights for alpha blending.

Learning Volume Rendering Priors. Our data-driven strategy uses ground truth meshes $\{S_h\}_{h=1}^{H}$ to learn the function f_w. For each shape S_h, we calculate its ground truth UDF f_{gt}^h, render A=100 depth images $\{D_a^h\}_{a=1}^{A}$ from randomly sampled view angles around it, and push the neural network learning f_w to render depth images $\{\tilde{D}_a^h\}_{a=1}^{A}$ to be as similar to $\{D_a^h\}_{a=1}^{A}$ as possible. During rendering, we leverage f_{gt}^h to provide ground truth unsigned distances at query q_n, which leaves the function f_w as the only learning target.

Specifically, along a ray V_k, we hierarchically sample N=128 queries $\{q_n\}$ to render a depth value through volume rendering using Eq.(1). We use the same sampling strategy introduced in NeUDF [25]. For each query q_n, we calculate its ground truth unsigned distance $u_n = f_{gt}^h(q_n)$ and the ground truth unsigned distances at its M=30 neighboring points $\{u'_m = f_{gt}^h(q'_m)\}_{m=1}^{M}$. Besides, we also use the sampling interval δ_m between q'_m and q'_{m+1} as another clue. We do not use the coordinates as a clue to pursue better generalization ability on unseen scenes with different coordinates. Therefore, we formulate the modeling of opaque density as,

$$\sigma_n = f_w(\{\delta_m\}_{m=1}^{M}, \{f_{gt}^h(q'_m)\}_{m=1}^{M}), q'_m \in NN(q_n). \tag{2}$$

Instead of handcrafted equations, we use a neural network with 6 layers to model the function f_w. It will learn a volume rendering prior which is a prior knowledge

of being a good renderer for UDFs. Obviously, it is more adaptive to different rays than handcrafted equations, and become more 3D aware with the flexibility of using a larger neighboring size. We train the network parameterized by θ_w by minimizing the rendering errors on depth images,

$$\min_{\theta_w} \sum_{h=1}^{H} \sum_{a=1}^{A} ||D_a^h - \tilde{D}_a^h||_2^2. \tag{3}$$

We do not involve RGB images in the learning of priors for better generalization ability. The improvements brought by our prior are shown in Fig. 5. We can accurately predict opaque densities with 3D awareness at arbitrary locations and robustness to unsigned distance changes.

Generalizing Volume Rendering Priors. We use the volume rendering prior represented by θ_w of f_w to estimate a UDF f_u from a set of RGB images $\{I_j\}_{j=1}^{J}$ of an unseen scene. We learn f_u by minimizing the rendering errors on RGB images.

Specifically, for a ray V_k, we hierarchically sample N=128 queries $\{q_n\}$ to render RGB values through volume rendering using Eq.(1). Similarly, we calculate the opaque density at each location q_n using Eq.(2), but using unsigned distances predicted by f_u as $\sigma_n = f_w(\{\delta_m\}_{m=1}^{M}, \{f_u(q'_m)_{m=1}^{M}\})$, not the ground truth ones when learning f_w, and keeping the parameters θ_w of f_w fixed during the generalizing procedure. Moreover, we use two neural networks to model the UDF f_u and the color funtion f_c parameterized by θ_u and θ_c, respectively. We jointly learn θ_u and θ_c by minimizing the errors between rendered RGB images $\{\tilde{I}_j\}_{j=1}^{J}$ and the ground truth below,

$$\mathcal{L}_{rgb} = \sum_{j=1}^{J} ||I_j - \tilde{I}_j||. \tag{4}$$

Our loss function is formulated with an additional Eikonal loss [47] $\mathcal{L}_e$ for regularization in the field,

$$\mathcal{L} = \mathcal{L}_{rgb} + \lambda \mathcal{L}_e, \tag{5}$$

where λ is a balance weight and set to 0.1 following previous work [41].

Using the learned parameters θ_u, we use the method introduced in [16] to extract the zero-level set of f_u as the surface.

Implementation Details. We implement our volume rendering priors network f_w as a 6-layer MLP with 256 hidden units and skip connections. Similar to previous work [26,41], the UDF function f_u is an 8-layer MLP with skip connections and the color function f_c is a 4-layer MLP with 256 hidden units. To control the smoothness of the UDF learning, similar as the trainable variance in NeuS [41], we utilize two parameter sets of f_w for early and later UDF inference stage, respectively. More implementation details can be found in the supplementary materials.

4 Experiments

In this section, we evaluate our method in surface reconstruction from multi-view RGB images. We report numerical and visual comparisons with the latest

methods of learning UDFs under the same experimental setting. We also report ablation studies to justify the effectiveness of our modules and the effect of key parameters.

4.1 Experiment Settings

Data for Learning Priors. We select one object from "car" category of ShapeNet dataset [5] and one from DeepFashion3D dataset [52] to form our training dataset for learning volume rendering priors. Note that there is no overlap between the selected object and the testing objects. Our ablation studies demonstrate that these two objects are sufficient to learn accurate volume rendering priors with good generalization capabilities across various shape categories. For each object, we first convert it into a normalized watertight mesh and then render 100 depth images with 600×600 resolution from uniformly distributed camera viewpoints on a unit sphere. Without additional annotation, we utilize the volume rendering priors pre-trained on these two shapes to report our results.

Datasets for Evaluations. We evaluate our methods on four datasets including DeepFashion3D dataset (DF3D) [52], DTU dataset [19], Replica dataset [35] and real-captured datasets. For DF3D dataset, we follow [26] and use the same 12 garments from different categories. For DTU dataset, we use the same 15 scenes that are widely used by previous studies. And we use all the 8 scenes in Replica dataset. We also report results on real scans from NeUDF [25] and the ones shot by ourselves.

Baselines. We compare our method with the state-of-the-art methods which use different differentiable renderers to reconstruct open surfaces, including NeuralUDF [26], NeUDF [25] and NeAT [29]. We also report the results of NeuS [41] and COLMAP [34] as baselines. Note that NeuralUDF uses additional patch loss [27] to fine-tune the resulted meshes, which is not the primary contribution of the differentiable renderer or used by other methods. Hence, for fair comparison, we report the results of NeuralUDF without fine-tuning across all datasets. However, we still perform additional experiments in the supplementary to show that our method, when getting fine-tuned using the patch loss, outperforms NeuralUDF under the same experimental conditions.

Metrics. For DTU dataset and DF3D dataset, we use Chamfer Distance (CD) as the metric. For Replica dataset, we report CD, Normal Consistency (N.C.) and F1-score following previous works [3, 49]. Moreover, we report the rendering errors in Tab. 1 using depth L1 distance, mask errors with cross entropy and L1 distance. The definitions of the metrics are provided in the supplementary materials.

Table 1: Numerical comparisons in all ShapeNet categories.

Methods	Depth-L1↓	Mask-Entropy↓	Mask-L1↓
NeuS-UDF [41]	3.46±1.49	2.67±1.27	2.27±1.09
NeUDF [25]	1.67±0.31	4.19±2.02	0.89±0.14
NeuralUDF [26]	1.28±0.28	30.65±7.59	0.61±0.12
NeuS-SDF [41]	0.97±0.67	**0.60±0.56**	0.52±0.47
Ours	**0.41±0.19**	0.70±0.63	**0.12±0.05**

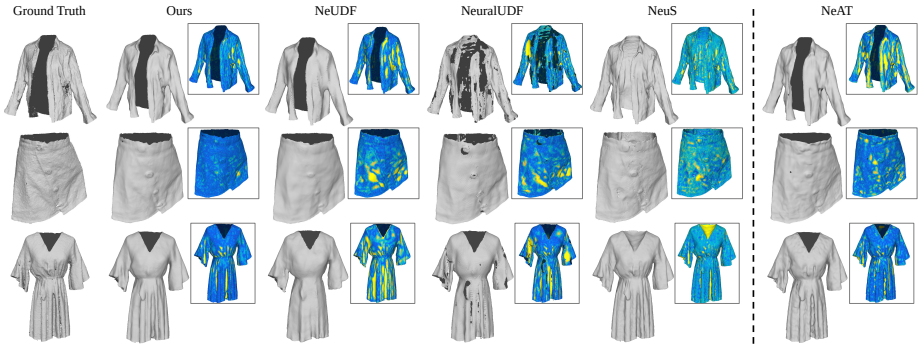


Fig. 6: Visual comparisons on open surface reconstructions with error maps on DeepFashion3D [52] dataset (NeAT uses additional mask supervision). The large reconstruction errors are shown in yellow on error maps.

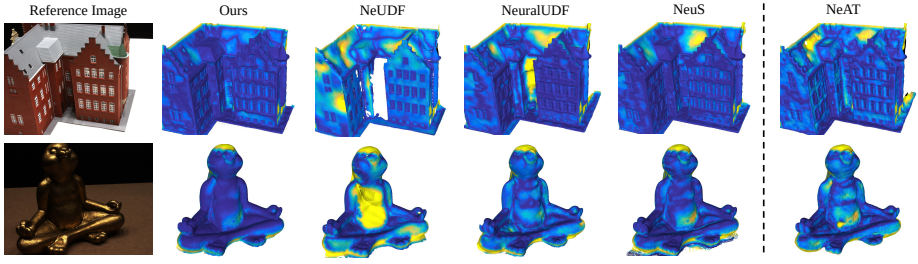


Fig. 7: Visual comparisons of error maps on DTU [19] dataset. The transition from blue to yellow indicates larger reconstruction errors.

4.2 Comparisons with the Latest

Results on ShapeNet. Tab. 1 reports numerical comparisons in the experiment in Fig. 2 in terms of 3 metrics. We report the averages and variances over all 55 shapes. We render depth images and mask images by forwarding ground truth unsigned distances at the same set of queries as other methods with the learned prior. We calculate the L1-error and cross entropy error between predicted images and ground truth ones. We achieve the best accuracy among all renderers for UDFs and SDFs.

Table 2: Quantitative evaluations on DF3D [52], DTU [19] and Replica [35] datasets. Note that NeAT uses mask supervision.

Datasets	DF3D	DTU	Replica		
Metrics	CD↓	CD↓	CD↓	N.C.↑	F-score↑
NeAT [29]	2.10	0.88	0.18	0.75	0.36
COLMAP [34]	3.10	1.36	0.23	0.46	0.43
NeuS [41]	4.36	0.87	0.07	0.88	0.69
NeuralUDF [26]	2.15	1.07	0.11	0.85	0.53
NeUDF [25]	2.01	1.58	0.28	0.78	0.31
Ours	**1.71**	**0.85**	**0.04**	**0.90**	**0.80**

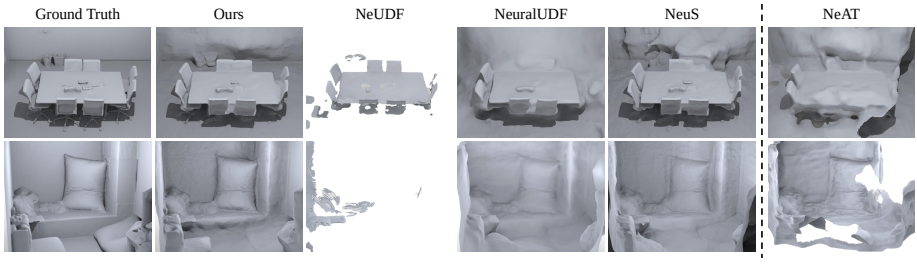


Fig. 8: Qualitative comparisons on Replica [35] dataset (NeAT uses additional mask supervision). Our method outperforms other methods on complex indoor scenes while other UDF-based methods struggle to recover complete and smooth surfaces.

Results on DF3D. We first report evaluations on DF3D (CD $\times 10^{-3}$). Numerical comparison in Tab. 2 indicates that our learned prior produces the lowest CD errors among all handcrafted renderers. The visual comparison in Fig. 6 details our superiority on reconstruction with error maps. We see that our prior helps the network to recover not only smoother surfaces at most areas but also sharper edges at wrinkles. We also outperform NeAT [29] which uses additional mask supervisions to learn local SDFs and reconstruct open surfaces. Please see our supplementary materials for evaluations on each scene.

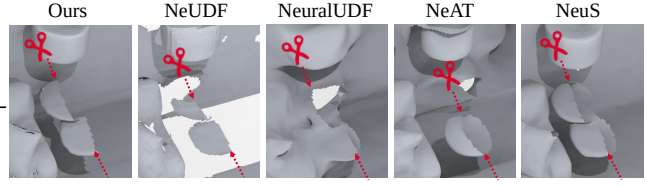


Fig. 9: Illustration of the capabilities of reconstructing single-layer geometries in indoor scenes.

Results on DTU. Tab. 2 reports our evaluation on DTU. Our reconstructions produce the lowest CD errors with our pre-trained prior. Although the shapes used to learn the prior are not related to any scenes in DTU, our prior comes from sets of queries in a local window on a ray, which are more general to unsigned dis-

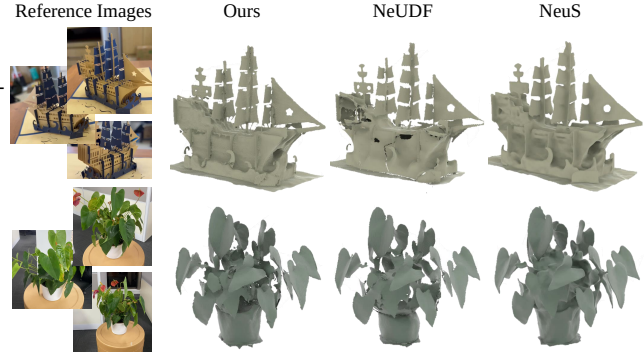


Fig. 10: Visualization of our real-captured scenes.

tances along a ray in unobserved scenes. Hence, our prior produces excellent generalization ability. Visual comparisons in Fig. 7 detail our reconstruction accuracy. Please see our supplementary materials for evaluations on each scene.

Results on Replica. We also evaluate our method in large-scale indoor scenes. Numerical valuations in Tab. 2 show that we produce much lower reconstruction errors than handcrafted renderers for UDFs, and even for SDFs. Visual comparisons in Fig. 8 show that our prior can recover geometry with higher accuracy, sharper edges, and much less artifacts. For objects that we can only observe from one side, our method is able to reconstruct it as a single surface, as illustrated in Fig. 9, which justifies the unsigned distance character. Please see our supplementary materials for evaluations on each scene.

Results on Real Scans. We further compare our method with the latest UDF reconstruction method NeUDF [25] on real scans in Fig. 1 and Fig. 10. We shot 4 video clips on 4 scenes with thin surfaces. The comparisons show that these challenging cases make NeUDF struggle to recover extremely thin surfaces like egg shells, resulting in incomplete and discontinuous surfaces. Comparing to the SDF learned by NeuS, our reconstruction with UDF produces much sharper structures. Similarly, we also produce more accurate and smoother surfaces than NeUDF on the real scans used in NeUDF, as shown in Fig. 11. Please watch our video for more details.

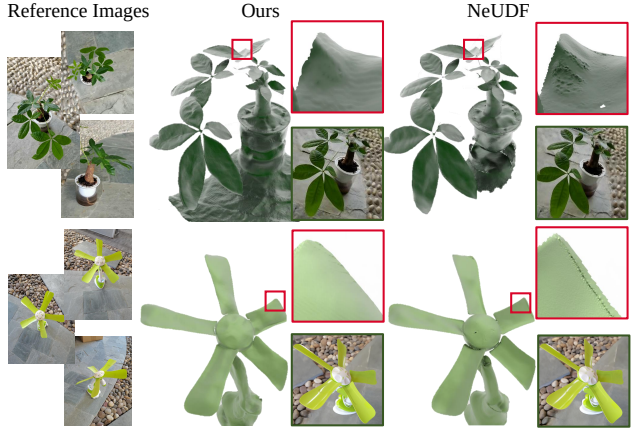


Fig. 11: Visualization of real scans used in NeUDF. The righttop and rightbottom part of each image are enlarged details and rendering views, respectively.

4.3 Ablation Studies and Analysis

Geometry Overfitting with Depth Supervision. We first report analysis on the capability of geometry reconstruction from depth images on a single shape, which highlights the performance of our prior over others without the performance of color modeling. We learn a UDF with our prior or other handcrafted renderers from multiple depth images. Visual comparison in Fig. 12 shows that handcrafted renderers do not recover geometry detail even in an overfitting experiment, while our learned prior can recover more accurate geometry than others.

Neighboring Size. We report the effect of window size in our volume rendering prior on DF3D and DTU datasets in Fig. 13 and Fig. 14a. We train different volume rendering priors with different window sizes including $\{1, 10, 20, 30, 40, 50\}$. With a small window, such as 1 and 10, the prior becomes very sensitive to unsigned distance changes, which produces holes on the surface. With larger window sizes, such as 50, the prior produces artifacts on the boundary. We find that a window covering 30 queries works well in our experiments.

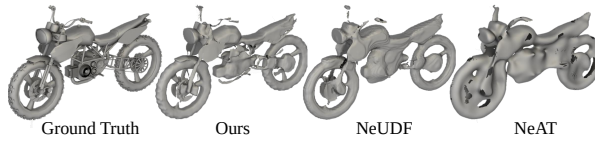


Fig. 12: Comparison of the ability of overfitting single complex object using ground truth depth supervision.

Shapes for Learning Priors. We explore the effect of the number of shapes used for learning priors in three more settings: (1) one shape from ShapeNet, (2) two shapes from ShapeNet and one from DF3D, (3) two shapes from ShapNet and another two from DF3D. The generalization results on DF3D and DTU datasets are presented in Fig. 14b. The prior learned from a single shape exhibits a severe underfitting on DF3D, while more than two shapes do not bring further improvements. We further show the simplicity and robustness of learning our prior from different sets of shapes in Tab. 4. Each set still contains one randomly sampled cloth and one shape from different ShapeNet classes, which does not affect our appealing performance. The reason is that we use lots of rays from different view angles to provide adequate knowledge of transforming unsigned distances to densities, which covers almost all unobserved situations in volume rendering for the UDF inference during testing.

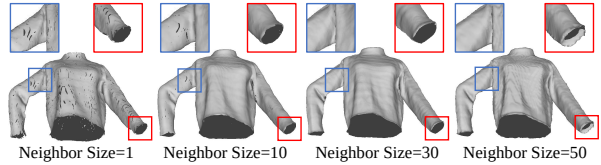
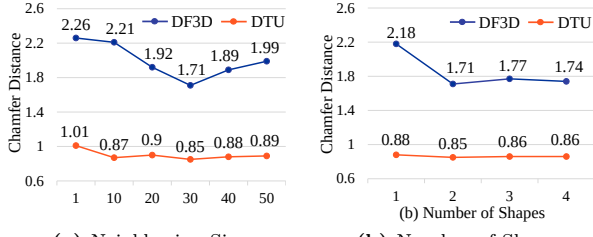


Fig. 13: Ablation study on the neighboring size.

Queries for Implicit Representations. We justify the superiority of using sampling interval along the ray as queries. We try to remove the interval or replace the interval using other alternative like coordinates. The degenerated results in the "Inputs" column in Tab. 3 indicate that the relative position represented by the interval generate better on unobserved scenes.



(a) Neighboring Size **(b)** Number of Shapes

Fig. 14: Ablation study on the neighboring size and number of shapes. The numerical results are averaged across all scenes in DTU and DF3D datasets.

Table 3: Ablation studies on prior variants. The numerical results are averaged across all objects in DF3D dataset.

Forms	Raw	Inputs		Supervisions		Inference
Variants	Ours	only UDF	UDF+Coor	RGB Sup.	RGBD Sup.	Fine-tune
CD-L1	**1.71**	1.87	2.92	Divergent	2.02	Divergent

Table 4: Choice of different training shapes

	Car+Cloth1	Chair+Cloth2	Sofa+Cloth3	Airplane+Cloth4	Bed+Cloth5
CD↓	1.71	1.78	1.72	1.74	1.74

Supervisions for Learning Priors. We further replace the supervisions of learning priors from depth images to RGB images or RGBD images, as reported in the "Supervisions" column in Tab. 3. Training with only RGB supervisions does not converge while the RGBD supervision severely degenerates the performance due to the aliasing of the color net. This indicates that the color affects the generalization ability of the prior a lot and is not suitable for learning priors for UDF rendering.

Fine-tuning Priors. Instead of using fixed parameters in the learned prior, we fine-tune the parameters of f_w using RGB images as supervision during testing, as reported in the "Inference" column in Tab. 3. We find that the optimization does not converge. This indicates that the prior has acquired sufficient generalization ability during training, requiring no further adjustments during testing.

Learning with SDF. We also try to use our method to learn a prior for SDF instead of UDF with the same setting. Comparison in Fig. 15 shows similar results between UDF and SDF priors. The superior results over NeuS demonstrates the effectiveness of our approach of learning volume rendering priors for both UDFs and SDFs.

Fig. 15: Ablation study on learning priors with SDF.

5 Conclusion

We introduce volume rendering priors for UDF inference from multi-view images through neural rendering. We show that using data-driven manner to learn the prior can recover more accurate geometry than handcrafted equations in differentiable renderers. We successfully learn a prior from depth images from few shapes using our novel neural network and learning scheme, robustly generalize the learned prior for UDFs inference from RGB images. We find that observing various unsigned distance variations during training and being 3D aware are the key to a prior with unbiasness, robustness, and scalability. Our extensive experiments and analysis on widely used benchmarks and real images justify our claims and demonstrate superiority over the state-of-the-art methods.

References

1. Arandjelović, R., Zisserman, A.: Nerf in detail: Learning to sample for view synthesis. arXiv preprint arXiv:2106.05264 (2021) 4
2. Azinović, D., Martin-Brualla, R., Goldman, D.B., Nießner, M., Thies, J.: Neural RGB-D surface reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6290–6301 (2022) 4
3. Azinović, D., Martin-Brualla, R., Goldman, D.B., Nießner, M., Thies, J.: Neural rgb-d surface reconstruction. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 6290–6301 (2022) 9
4. Chabra, R., Lenssen, J.E., Ilg, E., Schmidt, T., Straub, J., Lovegrove, S., Newcombe, R.A.: Deep local shapes: Learning local SDF priors for detailed 3D reconstruction. In: European Conference on Computer Vision. vol. 12374, pp. 608–625 (2020) 1
5. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F.: ShapeNet: An Information-Rich 3D Model Repository. Tech. Rep. arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago (2015) 2, 9
6. Chang, J.H.R., Chen, W.Y., Ranjan, A., Yi, K.M., Tuzel, O.: Pointersect: Neural rendering with cloud-ray intersection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8359–8369 (2023) 4
7. Chen, R., Han, S., Xu, J., Su, H.: Point-based multi-view stereo network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1538–1547 (2019) 4
8. Chen, Z., Tagliasacchi, A., Funkhouser, T., Zhang, H.: Neural dual contouring. ACM Transactions on Graphics (TOG) **41**(4), 1–13 (2022) 1
9. Chibane, J., Mir, A., Pons-Moll, G.: Neural unsigned distance fields for implicit function learning. arXiv **2010.13938** (2020) 1
10. Corona, E., Hodan, T., Vo, M., Moreno-Noguer, F., Sweeney, C., Newcombe, R., Ma, L.: Lisa: Learning implicit shape and appearance of hands. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20533–20543 (2022) 1
11. Curless, B., Levoy, M.: A volumetric method for building complex models from range images. In: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques. pp. 303–312 (1996) 4
12. Darmon, F., Bascle, B., Devaux, J.C., Monasse, P., Aubry, M.: Improving neural implicit surfaces geometry with patch warping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6260–6269 (2022) 4
13. Ding, Y., Yuan, W., Zhu, Q., Zhang, H., Liu, X., Wang, Y., Liu, X.: TransMVSNet: Global context-aware multi-view stereo network with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8585–8594 (2022) 4
14. Geng, C., Peng, S., Xu, Z., Bao, H., Zhou, X.: Learning neural volumetric representations of dynamic humans in minutes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8759–8770 (2023) 1
15. Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., Tan, P.: 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. pp. 2495–2504 (2020) 4

16. Guillard, B., Stella, F., Fua, P.: Meshudf: Fast and differentiable meshing of unsigned distance field networks. In: European Conference on Computer Vision. pp. 576–592. Springer (2022) 1, 5, 8
17. Guo, H., Peng, S., Lin, H., Wang, Q., Zhang, G., Bao, H., Zhou, X.: Neural 3d scene reconstruction with the manhattan-world assumption. In: IEEE Conference on Computer Vision and Pattern Recognition (2022) 4
18. Hou, F., Deng, J., Chen, X., Wang, W., He, Y.: Neudf: Learning unsigned distance fields from multi-view images for reconstructing non-watertight models. arXiv preprint arXiv:2303.15368 (2023) 2, 3, 4, 5
19. Jensen, R., Dahl, A., Vogiatzis, G., Tola, E., Aanæs, H.: Large scale multi-view stereopsis evaluation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 406–413 (2014) 9, 10
20. Jiang, C., Sud, A., Makadia, A., Huang, J., Nießner, M., Funkhouser, T., et al.: Local implicit grid representations for 3D scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6001–6010 (2020) 1
21. Kazhdan, M.M., Hoppe, H.: Screened poisson surface reconstruction. ACM Transactions Graphics **32**(3), 29:1–29:13 (2013) 4
22. Kong, X., Liu, S., Taher, M., Davison, A.J.: vmap: Vectorised object mapping for neural field slam. arXiv preprint arXiv:2302.01838 (2023) 4
23. Kurz, A., Neff, T., Lv, Z., Zollhöfer, M., Steinberger, M.: Adanerf: Adaptive sampling for real-time rendering of neural radiance fields. In: European Conference on Computer Vision. pp. 254–270. Springer (2022) 4
24. Li, Z., Müller, T., Evans, A., Taylor, R.H., Unberath, M., Liu, M.Y., Lin, C.H.: Neuralangelo: High-fidelity neural surface reconstruction. In: IEEE Conference on Computer Vision and Pattern Recognition (2023) 4
25. Liu, Y.T., Wang, L., Yang, J., Chen, W., Meng, X., Yang, B., Gao, L.: Neudf: Leaning neural unsigned distance fields with volume rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 237–247 (2023) 2, 3, 4, 5, 6, 7, 9, 10, 12
26. Long, X., Lin, C., Liu, L., Liu, Y., Wang, P., Theobalt, C., Komura, T., Wang, W.: Neuraludf: Learning unsigned distance fields for multi-view reconstruction of surfaces with arbitrary topologies. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20834–20843 (2023) 1, 2, 3, 4, 5, 6, 8, 9, 10
27. Long, X., Lin, C., Wang, P., Komura, T., Wang, W.: Sparseneus: Fast generalizable neural surface reconstruction from sparse views. In: European Conference on Computer Vision. pp. 210–227. Springer (2022) 9
28. Ma, B., Han, Z., Liu, Y.S., Zwicker, M.: Neural-Pull: Learning signed distance functions from point clouds by learning to pull space onto surfaces. CoRR **abs/2011.13495** (2020) 1
29. Meng, X., Chen, W., Yang, B.: Neat: Learning neural implicit surfaces with arbitrary topologies from multi-view images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 248–258 (2023) 2, 3, 4, 5, 9, 10, 11
30. Niemeyer, M., Mescheder, L., Oechsle, M., Geiger, A.: Differentiable volumetric rendering: Learning implicit 3D representations without 3D supervision. In: IEEE Conference on Computer Vision and Pattern Recognition (2020) 4
31. Oechsle, M., Niemeyer, M., Reiser, C., Mescheder, L., Strauss, T., Geiger, A.: Learning implicit surface light fields. In: 2020 International Conference on 3D Vision (3DV). pp. 452–462. IEEE (2020) 1

32. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: DeepSDF: Learning continuous signed distance functions for shape representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 165–174 (2019) 1
33. Rosu, R.A., Behnke, S.: Permutosdf: Fast multi-view reconstruction with implicit surfaces using permutohedral lattices. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 4
34. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4104–4113 (2016) 9, 10
35. Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J.J., Mur-Artal, R., Ren, C., Verma, S., et al.: The replica dataset: A digital replica of indoor spaces. arXiv preprint arXiv:1906.05797 (2019) 9, 10, 11
36. Takikawa, T., Litalien, J., Yin, K., Kreis, K., Loop, C., Nowrouzezahrai, D., Jacobson, A., McGuire, M., Fidler, S.: Neural geometric level of detail: Real-time rendering with implicit 3D shapes. In: IEEE Conference on Computer Vision and Pattern Recognition (2021) 1
37. Wang, H., Wang, J., Agapito, L.: Co-slam: Joint coordinate and sparse parametric encodings for neural real-time slam (2023) 4
38. Wang, J., Wang, P., Long, X., Theobalt, C., Komura, T., Liu, L., Wang, W.: NeuRIS: Neural reconstruction of indoor scenes using normal priors. In: European Conference on Computer Vision (2022) 4
39. Wang, J., Bleja, T., Agapito, L.: Go-surf: Neural feature grid optimization for fast, high-fidelity rgb-d surface reconstruction. In: International Conference on 3D Vision (2022) 4
40. Wang, L., Chen, W., Meng, X., Yang, B., Li, J., Gao, L., et al.: Hsdf: Hybrid sign and distance field for modeling surfaces with arbitrary topologies. Advances in Neural Information Processing Systems **35**, 32172–32185 (2022) 1
41. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. Advances in Neural Information Processing Systems **34** (2021) 2, 3, 4, 8, 9, 10
42. Wang, Y., Skorokhodov, I., Wonka, P.: HF-NeuS: Improved surface reconstruction using high-frequency details. In: Advances in Neural Information Processing Systems (2022) 4
43. Weilharter, R., Fraundorfer, F.: HighRes-MVSNet: A fast multi-view stereo network for dense 3D reconstruction from high-resolution images. IEEE Access **9**, 11306–11315 (2021) 4
44. Yan, J., Wei, Z., Yi, H., Ding, M., Zhang, R., Chen, Y., Wang, G., Tai, Y.W.: Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. In: European Conference on Computer Vision. pp. 674–689. Springer (2020) 4
45. Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: Mvsnet: Depth inference for unstructured multi-view stereo. European Conference on Computer Vision (2018) 4
46. Yariv, L., Gu, J., Kasten, Y., Lipman, Y.: Volume rendering of neural implicit surfaces. In: Advances in Neural Information Processing Systems (2021) 1, 4
47. Yariv, L., Kasten, Y., Moran, D., Galun, M., Atzmon, M., Ronen, B., Lipman, Y.: Multiview neural surface reconstruction by disentangling geometry and appearance. Advances in Neural Information Processing Systems **33** (2020) 8
48. Yu, Z., Gao, S.: 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. pp. 1949–1958 (2020) 4

49. Yu, Z., Peng, S., Niemeyer, M., Sattler, T., Geiger, A.: MonoSDF: Exploring monocular geometric cues for neural implicit surface reconstruction. ArXiv **abs/2022.00665** (2022) 4, 9
50. Zhao, D., Lichy, D., Perrin, P.N., Frahm, J.M., Sengupta, S.: Mvpsnet: Fast generalizable multi-view photometric stereo. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12525–12536 (2023) 4
51. Zhou, J., Ma, B., Liu, Y.S., Fang, Y., Han, Z.: Learning consistency-aware unsigned distance functions progressively from raw point clouds. In: Advances in Neural Information Processing Systems (NeurIPS) (2022) 1, 5
52. Zhu, H., Cao, Y., Jin, H., Chen, W., Du, D., Wang, Z., Cui, S., Han, X.: Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. pp. 512–530. Springer (2020) 9, 10
53. Zhu, Z., Peng, S., Larsson, V., Cui, Z., Oswald, M.R., Geiger, A., Pollefeys, M.: NICER-SLAM: neural implicit scene encoding for RGB SLAM. CoRR **abs/2302.03594** (2023) 4
54. Zhu, Z., Peng, S., Larsson, V., Xu, W., Bao, H., Cui, Z., Oswald, M.R., Pollefeys, M.: Nice-slam: Neural implicit scalable encoding for slam. In: IEEE Conference on Computer Vision and Pattern Recognition (2022) 4