Efficient Implicit SDF and Color Reconstruction via Shared Feature Field

Shuangkang Fang[1*], Dacheng Qi[1*], Weixin Xu[2], Yufeng Wang[3(✉)], Zehao Zhang[1], Xiaorong Zhang[1], Huayu Zhang[1], Zeqi Shao[1], and Wenrui Ding[3]

[1] School of Electronic Information Engineering, Beihang University, Beijing, China
{skfang,dc_qi,zhangzehao,zhangxiaorong,huayuzhang,zqshao}@buaa.edu.cn
[2] MEGVII Technology, Beijing, China
{xuweixin02}@megvii.com
[3] Institute of Unmanned System, Beihang University, Beijing, China
{wyfeng,ding}@buaa.edu.cn

Abstract. Recent advancements in neural implicit 3D representations have enabled simultaneous surface reconstruction and novel view synthesis using only 2D RGB images. However, these methods often struggle with textureless and minimally visible areas. In this study, we introduce a simple yet effective encoder-decoder framework that encodes positional and viewpoint coordinates into a shared feature field (SFF). This feature field is then decoded into an implicit signed distance field (SDF) and a color field. By employing a weight-sharing encoder, we enhance the joint optimization of the SDF and color field, enabling better utilization of the limited information in the scene. Additionally, we incorporate a periodic sine function as an activation function, eliminating the need for a positional encoding layer and significantly reducing rippling artifacts on surfaces. Empirical results demonstrate that our method more effectively reconstructs textureless and minimally visible surfaces, synthesizes higher-quality novel views, and achieves superior multi-view reconstruction with fewer input images.

Keywords: 3D Multi-view Representation · Textureless · signed distance field.

1 Introduction

3D scene reconstruction from multiple 2D images is a fundamental challenge in both computer graphics and computer vision [14, 37, 48, 50, 57]. Recent advances in neural implicit representations have demonstrated significant potential in reconstructing appearance and geometry [13, 16, 32, 56], as well as in synthesizing novel views [8, 15, 17, 37, 39]. By leveraging rendering methods, frameworks that represent implicit surfaces through coordinate-based neural networks enable the

*Equal contribution; ✉ Corresponding author

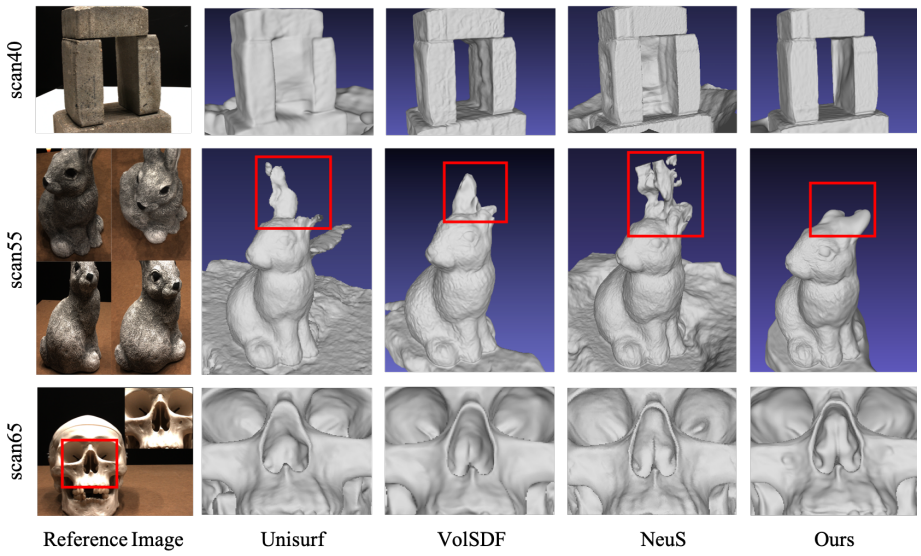

Fig. 1: Qualitative results of typical examples. Using roughly half of the training images in the DTU MVS dataset [25], our method outperforms Unisurf [44], VolSDF [63], and NeuS [56] in learning more accurate and complete surfaces in textureless regions with complex topology (e.g., scan40 and scan65). In scarcely visible regions, such as the ears of the rabbit (scan55), our method generates a cleaner surface without extraneous or missing parts.

conversion of 3D representations into 2D views. These frameworks are differentiable and rely solely on 2D images for ground truth.

The effectiveness of reconstruction hinges on the choice of implicit representations, rendering techniques, density modeling, and the co-optimization of various fields. Recent studies have successfully combined signed distance fields (SDF) [42, 56, 63, 64] or occupancy fields [44] with surface [42, 64] or volume [44, 56, 63] rendering techniques, resulting in notably improved performance. These methods co-optimize different fields by either directly determining radiance on surfaces [42, 64] or by transforming learned implicit fields into local transparency functions for volume rendering [44, 56, 63], capturing both the geometry and appearance of solid and non-transparent 3D scenes with high fidelity. Moreover, Gaussian-Splatting [28] employs anisotropic 3D Gaussian primitives to facilitate real-time reconstruction for a scene. Based on this, 2DGS [23] adopts anisotropic 2D Gaussian primitives to streamline the extraction of surface information. However, challenges persist in accurately reconstructing textureless and rarely visible surfaces. As illustrated in Fig. 1, in scenes with complex topology or textureless regions that lack pixel-wise object masks for training supervision, these methods struggle to fully recover the scene.

In this paper, we introduce SFF to address challenges associated with textureless and rarely visible regions. The core concept is that different neural fields

share a common feature from an encoder, which reduces ambiguity during the co-optimization of these fields by enriching each with supplementary information from others. Specifically, for each spatial position (x, y, z), only a single forward pass is needed. Unlike prior approaches that process the same spatial positions through different fields and link these fields with intermediate features, our method facilitates direct information sharing across fields. Each spatial position is encoded once and then decoded into the specific values of different fields using shallow decoders, some with as few as one linear layer.

Additionally, we observe that optimizing surface reconstruction often results in corrugated artifacts. Experiments indicate that these artifacts might be caused by high-frequency components introduced by positional encoding, which are ill-suited for this framework. To address this, we propose eliminating the positional encoding layer and using a periodic sine function as the activation in the encoder. This approach reduces corrugated artifacts and enhances surface smoothness.

In summary, our key contributions are:

- We present an end-to-end Encoder-Decoder framework that conducts a single forward pass of spatial positions to encode diverse representations, which are then decoded into various fields. This approach enhances surface reconstruction and novel view synthesis for 3D scenes with complex topology, particularly in textureless and rarely visible regions.
- We introduce a hybrid Sinus-ReLU activation function that supersedes the positional encoding layer, effectively diminishing mesh artifacts while preserving the quality of novel view synthesis.
- Experimental results demonstrate that the proposed method more effectively utilizes the limited information available in a scene, achieving high-quality surface and appearance reconstruction.

2 Related Works

Multi-view 3D reconstruction methodologies have undergone substantial evolution, shifting from traditional techniques to sophisticated, learning-based approaches. Classical methods primarily focus on extracting and reconstructing information from feature points [5, 7, 18, 19, 34, 51, 54] or voxel grids [1, 6, 12, 29, 31, 52]. Feature-based techniques involve matching features across adjacent views to generate depth maps for each pixel, subsequently requiring extensive post-processing to achieve watertight surfaces. This includes depth information fusion [11, 34] and mesh reconstruction processes, such as Poisson Surface Reconstruction [27]. However, this pipeline is complex and prone to accumulating errors, particularly in scenarios with non-Lambertian or textureless surfaces, which can degrade reconstruction quality. Conversely, voxel-based methods directly render complete surfaces from a 3D voxel grid but are constrained by low resolutions due to high memory demands. With advancements in neural networks, learning-based methods have been introduced to optimize the intermediate stages of classical pipelines [30,46,49], such as improving feature matching [30], and developing end-to-end trainable systems that derive surfaces from

depth supervision [24, 60, 61]. Recent innovations leverage implicit representations and rendering techniques, enabling these methods to reconstruct watertight surfaces and synthesize novel views based solely on 2D image supervision, with [42, 64] or without [44, 56, 63] the need for pixel-wise object masks.

3D Implicit Representation and Differentiable Rendering. Implicit 3D representations emerge as a promising alternative due to their inherently continuous nature, enabling the representation of 3D scenes without the need for discretization [3, 35, 36, 41, 43, 45, 47]. These representations also offer the advantage of a smaller memory footprint compared to the substantial memory requirements of voxel-based methods. Implicit representations may take the form of signed distance fields [45], occupancy fields [10, 35], or other signed fields [4]. When combined with differentiable rendering, these representations can encode both the 3D appearance and geometry into 2D images, allowing for surface reconstruction under solely 2D supervision. Differentiable rendering techniques can generally be classified into two categories based on the radiance calculation method: surface rendering and volume rendering. Surface rendering techniques [42, 64] apply rendering functions directly to object surfaces to compute radiance but typically require pixel-wise masks for learning precise implicit representations. Conversely, volume rendering, as exemplified by NeRF [37], employs a radiance field's alpha composition along a ray to synthesize photo-realistic images. Recent multi-view 3D reconstruction methods [44, 56, 63] adopt this volume rendering approach, innovatively transforming the implicit field to model point density along a ray, thus achieving high-fidelity reconstruction of appearance and geometry without the need for mask supervision. Despite these advancements, these methods still face challenges in textureless or infrequently visible regions due to inherent ambiguities. Unlike these approaches, our method, SFF, successfully reproduces high-quality surfaces and novel views in such challenging conditions. We demonstrate that using a weight-shared encoder-decoder framework combined with a mixed Sinus-ReLU activation function effectively reduces ambiguities in textureless areas while preserving the integrity of the implicit field even in rarely observed regions.

Point-based Rendering. Recent advancements in point-based 3D rendering [9, 22, 23, 28, 33, 58, 59] demonstrate significant improvements in rendering efficiency. Notably, NPBG [2] leverages a convolutional neural network to generate RGB images from rasterized images. In contrast, 3DGS [28] employs anisotropic 3D Gaussian primitives to facilitate real-time reconstruction. Building on these methods, Relightable3DGS [20] introduces physically-based differentiable rendering to each 3D Gaussian point, incorporating lighting information to enable scene relighting. GaussianShader [26] simplifies the shading function to improve the performance of 3DGS on reflective surfaces. For surface reconstruction, NeuSG [9] utilizes an implicit SDF network to delineate surfaces from 3DGS. SuGaR [22] incorporates a custom regularization term to better align Gaussian primitives with the scene's surface. Additionally, 2DGS [23] adopts anisotropic 2D Gaussian primitives to streamline the extraction of surface information from 3DGS. Although point-based rendering methods are efficient in creating realistic

scenes, their high demand for RGB images is notable. In contrast, SFF achieves comparable 3D scene generation with fewer RGB images.

3 Method

In this section, we initially discuss the background encompassing the implicit field, volume rendering, and density modeling of the implicit field. Subsequently, we offer a detailed description of SFF, including the Encoder-Decoder paradigm and the mixed Sinus-ReLU activation function paradigm. An overview of our approach is illustrated in Fig. 2.

3.1 Background

Our goal is to reconstruct the geometry $\boldsymbol{G}$ and the appearance $\boldsymbol{A}$ of 3D objects from a set of 2D RGB images $\{\boldsymbol{I_j}\}$ with accurate camera poses. We employ the zero-level set of an implicit signed distance field (SDF), defined by $\theta \in \mathbb{R}^m$, to represent the geometry. The appearance is recovered through volume rendering, utilizing density modeling of the SDF and a color field, defined by $\gamma \in \mathbb{R}^n$. In this section, we present a brief background of the aforementioned ideas, for solid and non-transparent objects.

Signed Distance Field and Color Field. Each spatial position $\boldsymbol{x} \in \mathbb{R}^3$ at a viewing direction $\boldsymbol{v} \in \mathbb{R}^3$ is mapped to its signed distance to the object by $f(\boldsymbol{x};\theta) \rightarrow \mathbb{R}$, and to its color by $f(\boldsymbol{x},\boldsymbol{v};\gamma) \rightarrow \mathbb{R}^3$. These mappings are facilitated by a neural Multi-Layer Perceptron (MLP). The geometry of the object is therefore represented by the zero-level set of the SDF, which can be given by:

$$\boldsymbol{G} = \{\boldsymbol{x} \in \mathbb{R}^3 \mid f(\boldsymbol{x};\theta) = 0\}, \tag{1}$$

and the color of each position is given by:

$$c(\boldsymbol{x},\boldsymbol{v}) = f(\boldsymbol{x},\boldsymbol{v};\gamma). \tag{2}$$

Volume Rendering and Density Modeling. Following NeRF [37], along a ray $\boldsymbol{r}(t) = \{\boldsymbol{o} + t\boldsymbol{v} \mid t \geq 0\}$ through a pixel in a 2D image, where $\boldsymbol{o}$ is the center of the camera and $\boldsymbol{v}$ is the viewing direction, the color of this pixel $C(\boldsymbol{r})$ can be accumulated by:

$$C(\boldsymbol{r}) = \int_0^{+\infty} T(t)\sigma(\boldsymbol{r}(t))c(\boldsymbol{r}(t),\boldsymbol{v})\mathrm{d}t, \tag{3}$$

where $T(t) = \exp(-\int_0^t \sigma(\boldsymbol{r}(s))\mathrm{d}s$ denotes the accumulated transmittance along the ray and σ is the volume density. The solution to our goal has now become finding a proper method to model the volume density based on the SDF. In this paper, we adopt the method used in NeuS [56], where an opaque density function ρ is proposed as the counterpart of the volume rendering, given by:

$$\rho(t) = \max\left(\frac{-\frac{d\Phi_s}{\mathrm{d}t}(f(\boldsymbol{r}(t)))}{\Phi_s(f(\boldsymbol{r}(t)))}, 0\right), \tag{4}$$

where f is the SDF function, $\Phi_s(x) = (1+\mathrm{e}^{-sx})^{-1}$ is the Sigmoid function whose derivative $\phi_s(x)$ has a standard deviation of $1/s$. In NeuS, s is a trainable parameter, and $1/s$ approaches zero as the MLP converges during training. Finally, the color of a pixel in Eq. (3) can be rewritten using numerical quadrature based on Eq. (4) as:

$$C(\boldsymbol{r}) = \sum_{i=1}^{n} T_i \alpha_i c_i, \tag{5}$$

where $T_i = \prod_{j=1}^{i-1}(1-\alpha_j)$ is the discrete accumulated transmittance, and α_i is discrete opacity density, given by:

$$\alpha_i = \max\left(\frac{\Phi_{\mathrm{s}}(f(\boldsymbol{r}(t_i))) - \Phi_{\mathrm{s}}(f(\boldsymbol{r}(t_{i+1})))}{\Phi_{\mathrm{s}}(f(\boldsymbol{r}(t_i)))}, 0\right), \tag{6}$$

we refer readers to [56] for more details. Under this setting, we are able to co-optimize the SDF and color field by defining loss functions between the rendered images and the ground truth, which will be described in the next section.

Recent volume rendering-based works [44, 56, 63] pay lots of attention to the selection of implicit field and the density modeling, while they share a similar networks paradigm, where different MLPs are designed for different fields, and therefore each spatial position requires several forward passes. Besides, features extracted from one MLP will be fed to another, which may introduce ambiguity among different MLPs. For example, in [44, 56, 63], 2 MLPs, MLP_i with parameters θ_i and MLP_c with parameters θ_c represent the implicit SDF field and the color field respectively. Besides the implicit function value O_i, the features $\boldsymbol{F}$ and gradients ∇ output from one forward pass of the spatial positions $\boldsymbol{x}$ in the implicit field MLP are concatenated with the spatial positions again and the view direction $\boldsymbol{v}$, and then fed to the color field MLP for another forward pass to output the color c of each position. This paradigm, as given in Eq. (7) and Eq. (8), lacks regularization among different MLPs, and intuitively, it needs to match the input positions of one MLP to the features resulting from another forward pass of the same positions in another MLP.

$$O_i, \boldsymbol{F} = MLP_i(\boldsymbol{x}; \theta_i) \tag{7}$$

$$c = MLP_c(\boldsymbol{x}, \boldsymbol{v}, \nabla_{\boldsymbol{x}} O_i, \boldsymbol{F}; \theta_c). \tag{8}$$

3.2 SFF

In this section, we present SFF for multi-view reconstruction. We unite the multiple MLPs paradigm into an Encoder-Decoder paradigm, accompanied by a mixed Sinus-ReLU activation function paradigm.

Encoder-Decoder paradigm. Our motivation arises from observing that each MLP in the common paradigm serves both as an encoder and a decoder. In this

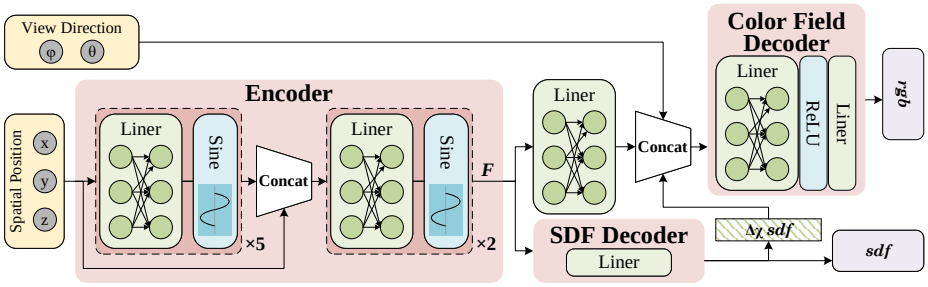

Fig. 2: Illustration of SFF. The proposed Encoder-Decoder paradigm encodes the spatial positions as a general feature in only one forward pass and decodes it into different fields.

setup, both the SDF and the radiance field employ separate MLPs with positional encoding, leading to ambiguity in matching features to positions. While such models fit the training views well, they often struggle to accurately represent the underlying geometric structure. To address this issue, we propose a new paradigm: a weight-shared encoder-decoder model. This model uses a single forward pass with spatial positions to obtain a general feature, which is then shared across multiple decoders. This approach provides greater regularization and reduces ambiguity in matching spatial positions to features across different forward passes.

To be specific, for a given position $\boldsymbol{x} \in \mathbb{R}^3$, an encoder E first encodes it into a general feature $\boldsymbol{F}$. The feature is then fed into 2 decoders, D_s and D_c, which represent the SDF and the color field, respectively. A forward pass in this encoder-decoder pipeline is given by:

$$\boldsymbol{F} = E(\boldsymbol{x}; \theta_E) \tag{9}$$

$$sdf(\boldsymbol{x}) = D_s(\boldsymbol{F}; \theta_{D_s}) \tag{10}$$

$$c(\boldsymbol{x}) = D_c(f(\boldsymbol{F}), \boldsymbol{v}, \nabla_{\boldsymbol{x}} sdf; \theta_{D_c}), \tag{11}$$

where θ_E, $\theta_{D_s}, \theta_{D_c}$ represent the learnable parameters of the encoder E, decoder D_s and D_c, respectively. $sdf(\boldsymbol{x})$ is the SDF value, and $c(\boldsymbol{x})$ is the RGB color value of point $\boldsymbol{x}$. f in Eq. (11) is a linear layer for the transformation of semantics. $\boldsymbol{v}$ is the viewing direction and $\nabla_{\boldsymbol{x}} sdf$ represents the normal vector of SDF.

Compared to the paradigm introduced in Sec. 3.1, where the color field is conditioned on $[\boldsymbol{x}, \boldsymbol{v}, \nabla_{\boldsymbol{x}} sdf, \boldsymbol{F}_{sdf}]$, we allow the SDF and color field to share the same feature $\boldsymbol{F}$ from the encoder and save the spatial position from another forward pass.

Mixed Sinus-ReLU Activation Function Paradigm. Following NeRF [37], a positional encoding layer for capturing high-frequency geometry and texture is applied to the spatial positions and the viewing direction before they are fed into the encoder and the decoder. However, we empirically find that high-frequency

positional encoding of the spatial position brings corrugated artifacts on the mesh, while low-frequency positional encoding results in relatively low-quality novel views.

To mitigate the corrugated artifacts and synthesize high-quality novel views, we remove the positional encoding layer and introduce a mixed Sinus-ReLU activation function paradigm for the encoder and decoders in our pipeline. Specifically, we use the sinus function as the activation function for the encoder, and ReLU [40] for the decoders. We follow SIREN [53] to set the sinus function as:

$$\mathrm{Act}(\boldsymbol{x}) = \sin(\omega_0 \cdot \boldsymbol{x}), \tag{12}$$

where ω_0 is a trainable parameter, initially set to 3. Network weights are initialized using a uniform distribution $W \sim U(-\sqrt{6/ic}/\omega_0, \sqrt{6/ic}/\omega_0)$, except for the first layer, which is initialized as $W \sim U(-\sqrt{1/ic}, \sqrt{1/ic})$, where ic represents the number of input channels for each layer.

Experimental evidence indicates that position-dependent functions of densely sampled spatial points need higher-frequency variations. Sparse views, however, lack adequate supervision to model these functions, causing frequency-related artifacts. Our mixed Sinus-ReLU activation function paradigm allows the network to capture frequency information effectively, producing smooth meshes and high-quality novel views through the decoders.

3.3 Training

Loss Function. Given a set of 2D images with accurate camera poses as ground truth, we optimize the following loss function during training:

$$\mathcal{L} = \mathcal{L}_{ren} + \lambda \mathcal{L}_{reg}, \tag{13}$$

$\mathcal{L}_{ren}$ is the $L1$ loss between the rendering color C_{ren} of the pixels in a ray batch of size B and the corresponding ground truth C_{gt}, given by:

$$\mathcal{L}_{ren} = \frac{1}{\mathrm{B}} \sum_{k} \mathrm{L1}(C_{ren}, C_{gt}). \tag{14}$$

Following [21], an Eikonal term $\mathcal{L}_{reg}$ is adopted to regularize the SDF value of each sampled point on the ray batch. Assuming N points are sampled along a ray, the $\mathcal{L}_{reg}$ is given by:

$$\mathcal{L}_{reg} = \frac{1}{NB} \sum_{i,j} (\|\nabla_{\boldsymbol{x_{i,j}}} sdf(\boldsymbol{x_{i,j}})\| - 1)^2, \tag{15}$$

where $\boldsymbol{x_{i,j}}$ is the j^{th} sampled point on the i^{th} ray.

Sampling. In our approach, we adopt the hierarchical sampling strategy from NeuS [56], which involves an initial coarse sampling followed by a finer one, both

conducted within a single network. This differs from the hierarchical sampling methodology used in NeRF [37], where separate stages are typically employed. During the coarse sampling phase, the network remains untrained, and the standard deviation of the derivative of the Sigmoid function in Eq. (4), denoted as $1/s$, is set to a high, fixed value. Using this setting, we run the network to generate a probability distribution from samples obtained with $1/s$. This probability distribution is then used for the fine sampling stage, and then the network and the standard deviation are switched to be learnable for the training stage.

4 Experiments

In this section, we evaluate our method on the 15 common scenes from the DTU MVS dataset [25] and compare the quantitative results, including the Chamfer Distance [35] for meshes and the PSNR for novel views, with related baselines. We show that our method is better at dealing with objects with complicated topology in textureless and rarely visible regions with sparse training images.

4.1 Datasets

DTU MVS. The DTU MVS dataset [25] features various scenes of small models captured by an industrial robot arm in a dedicated studio, presenting challenges due to non-Lambertian surfaces and complex geometries. Each scene comprises either 49 or 64 images at a resolution of 1200×1600, accompanied by precise camera poses. We focus on 15 commonly analyzed scenes used in [42, 44, 56, 63, 64] from the DTU dataset. To demonstrate our method's efficacy with limited visibility, we only randomly select 20/30 images from each scene for training.
Blended MVS. The Blended MVS dataset [62] is an extensive MVS dataset. Unlike the DTU MVS dataset, it includes ambient lighting data, posing additional challenges for reconstructing dimly lit areas. As demonstrated in [38, 44, 55, 56, 63], each scene in the dataset contains 31 to 143 images and corresponding masks, all downscaled to a relatively low resolution of 768×576.

4.2 Baselines

For comparisons, we use the source code released by the following baselines to conduct experiments on the same dataset as our method, which means that we also decrease the number of training images from 49/64 to 20/30 for these methods. Except for the number of training images, we follow their settings to train on each scene and report the Chamfer Distance and PSNR.

- **COLMAP** [51] is a general-purpose Structure-from-Motion (SfM) and Multi-View Stereo (MVS) pipeline with a graphical and command-line interface. COLMAP estimates the depth of the scene according to the given image to obtain a dense point cloud for the scene. With the help of Poisson Surface Reconstruction [27], a mesh of the scene can be generated.

– **NeRF** [37] also provides an ability to extract mesh since the density is related to the transparency of a point in space. We define a density threshold of 30 to generate the mesh and use NeRF to synthesize novel views as well.
– **Unisurf** [44] models the 3D scenes with an implicit occupancy field and gradually learns to sample points near the surface. It co-optimizes the occupancy field and the color field to synthesize novel views.
– **VolSDF** [63] models the density from the SDF based on the cumulative distribution function of the Laplace distribution, with a zero mean and a learnable scale. It uses an adaptive sampling method to sample spatial points.
– **NeuS** [56] models the density from the SDF with a derivative of the Sigmoid function and adopts a hierarchical sampling method but with only one network. Unisurf, VolSDF, and NeuS use a similar network paradigm, as mentioned in Sec. 3.1.
– **2DGS** [23] is an explicit point-based rendering method based on the principles of 3DGS. It leverages ray-splat intersection and rasterization, utilizing 2D Gaussian primitives for enhanced surface approximation.
– **SuGaR** [22] employs a two-stage method leveraging 3D Gaussian primitives for mesh rendering. Initially, it utilizes 3DGS for a preliminary reconstruction. Subsequently, SuGaR incorporates a tailored regularization term that promotes alignment of the Gaussian primitives with the scene's surface, facilitating concurrent training of both the primitives and the mesh.

4.3 Implementation Details

As shown in Fig. 2, the encoder is modeled by an 8-layer MLP with a hidden size of 256, whose activation function is the sinus function. The decoder for the SDF is a one-layer MLP, which directly outputs the SDF value, and the decoder for the color field is a 2-layer MLP with a hidden size of 128 whose output is the RGB color. The activation function of the first layer of the color decoder is ReLU [40]. We assume that the foreground of the scenes is in a unit sphere and we sample 64 points for both the coarse and fine sampling procedures. For the background outside the sphere we use NeRF++ [65] as the rendering method and sample 32 points along each ray. The initialization method is described in Sec. 3.2. Each scene is trained on an NVIDIA A100 GPU for 1000k iterations.

4.4 Comparisons

Quantitative Comparison. Similar to [44, 56, 63, 64], we report the Chamfer Distance [35] between the reconstructed surfaces and the ground truths provided by the DTU MVS dataset to evaluate the reconstruction quality. We also report the PSNR on the testing set to evaluate the quality of the novel views. As shown in Tab. 1, our method outperforms the compared methods in mean Chamfer Distance and mean PSNR. Although in terms of the mean Chamfer Distance, our method is better than NeuS, we verify in the following part that in scenes that we focus on in this paper, those with complicated topology as well as textureless and rarely visible regions, our method performs much better.

Table 1: Quantitative results on the DTU dataset. The proposed method achieves better performance on the PSNR and Chamfer Distance (CD) metrics.

	ScanID	24	37	40	55	63	65	69	83	97	105	106	110	114	118	122	mean
CD	colmap	**1.14**	2.21	1.03	1.37	1.89	1.78	1.08	3.15	1.46	2.31	1.33	1.80	0.46	0.73	1.21	1.53
	NeRF	1.73	1.87	1.41	0.92	1.40	0.97	1.14	1.45	1.66	1.02	0.72	1.89	0.42	0.71	0.60	1.19
	VolSDF	1.72	1.94	**0.94**	0.60	1.25	0.81	1.02	1.45	1.36	0.92	0.76	2.06	0.46	0.76	0.68	1.12
	UniSurf	1.71	1.89	2.48	1.09	1.40	0.94	0.88	1.42	1.35	0.95	0.69	1.90	0.55	0.74	0.64	1.24
	NeuS	1.41	**1.24**	1.40	0.45	**1.17**	0.84	**0.67**	1.48	**1.15**	**0.82**	0.59	**1.32**	0.38	0.51	**0.55**	0.932
	Ours	**1.23**	1.36	1.13	**0.44**	1.22	**0.71**	0.72	**1.41**	1.29	0.92	**0.53**	1.56	**0.36**	**0.50**	0.56	**0.929**
PSNR	NeRF	22.06	20.59	21.61	24.65	24.53	24.84	25.10	27.06	21.46	24.90	29.98	28.86	27.72	32.21	31.96	25.84
	VolSDF	21.12	20.10	21.25	23.32	23.92	25.84	25.87	27.18	23.63	26.27	29.46	29.37	27.82	31.94	30.88	25.87
	UniSurf	22.45	20.72	19.82	24.35	21.69	26.38	26.00	26.92	24.80	25.23	31.09	30.38	28.30	32.98	32.49	26.24
	NeuS	23.37	23.71	24.97	30.09	29.77	**31.54**	**31.47**	31.35	29.66	30.76	**36.80**	**35.40**	**33.94**	**38.75**	**37.73**	31.29
	Ours	**26.74**	**25.83**	**26.99**	**30.13**	**30.40**	31.26	29.86	**32.16**	**29.89**	**30.94**	36.25	34.90	33.53	38.49	37.53	**31.49**

Qualitative Comparison. Under the setting of fewer training images, the reconstruction quality of Unisurf and VolSDF greatly degrades. COLMAP recovers detailed but incomplete surfaces because of the trimming. Similarly, incompleteness and lots of noise can be found in the reconstructed scene recovered by NeRF. Besides, the reconstruction quality of NeRF depends on the choice of the threshold of the density value. NeuS performs well in most of the scenes, while in scan40 and scan65 where the foreground object is textureless and with complicated topology, it is lost in ambiguous regions and misses some important details. On the contrary, our method succeeds in recovering the inner surface (scan40 in Fig. 1) and details like the windows on the building in scan24 (Fig. 3), even with less than half of the images in each scene for training. One more example is that, as shown in Fig. 1 in scan65, where a textureless skull with complicated topology and details on the surface is presented, our method accurately reconstructs the cartilage in the nasal cavity, and two tiny holes below the eye socket as well, while the other methods either fail to learn this structure (Unisurf and VolSDF), or reconstruct it incompletely (NeuS).

We also evaluate it against the recent Gaussian-splatting-based (GS-based) method, which shows great potential in surface reconstruction tasks. The results, as shown in Fig. 4, indicate that the GS-based method frequently produces discontinuous surfaces, characterized by multiple holes. This is primarily due to the explicit structure of the GS method, which struggles to optimize spatially continuous representations under sparse viewpoints. In contrast, our approach achieves superior surface reconstruction by jointly optimizing the SDF and radiance field through feature sharing, thereby extracting more comprehensive information from sparse scene views.

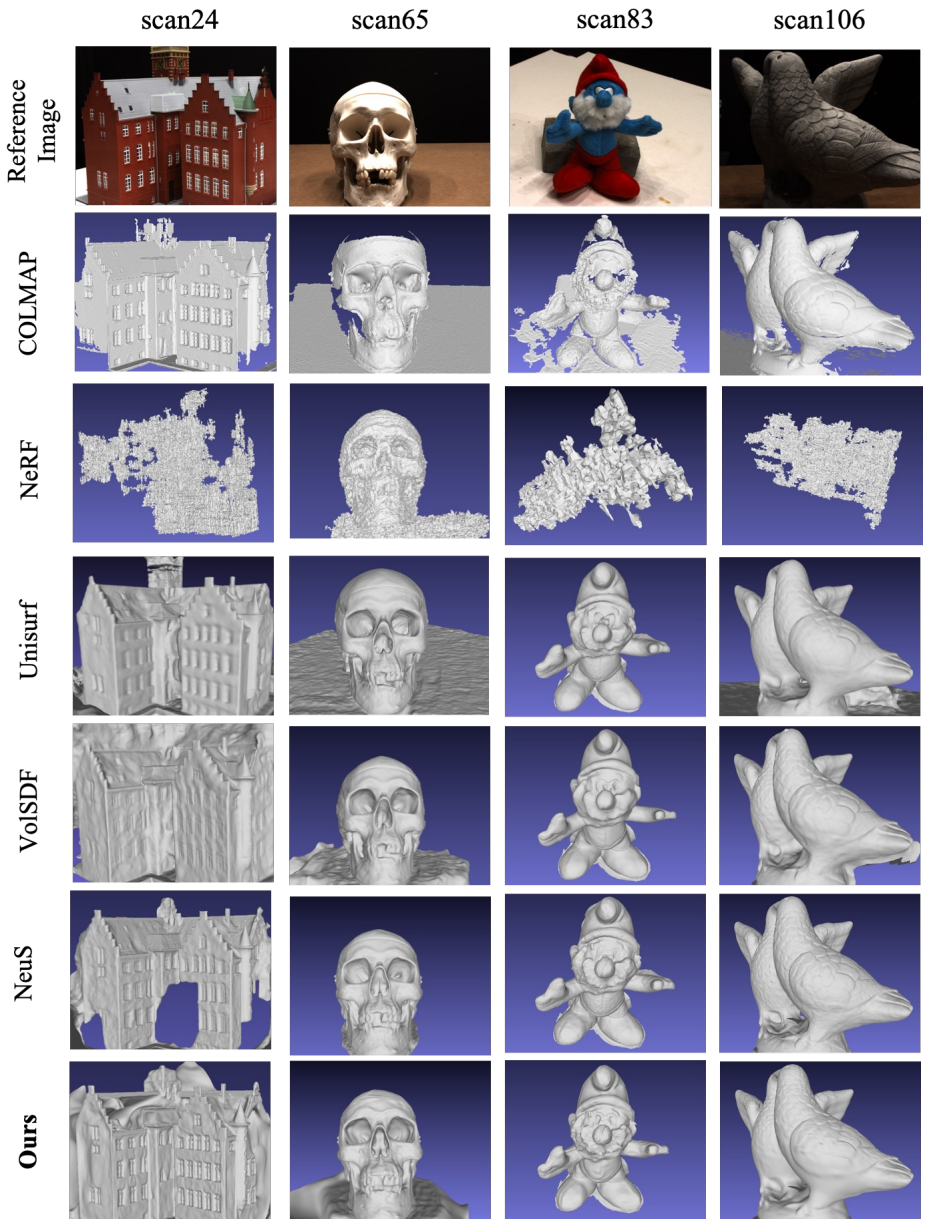

Fig. 3: Qualitative results. All of the scenes are trained under the setting in Sec. 4.1, where half the images are randomly sampled from the original training dataset for each scan. Compared to COLMAP [51], NeRF [37], Unisurf [44], VolSDF [63] and NeuS [56], our method learns more complete surfaces with more details.

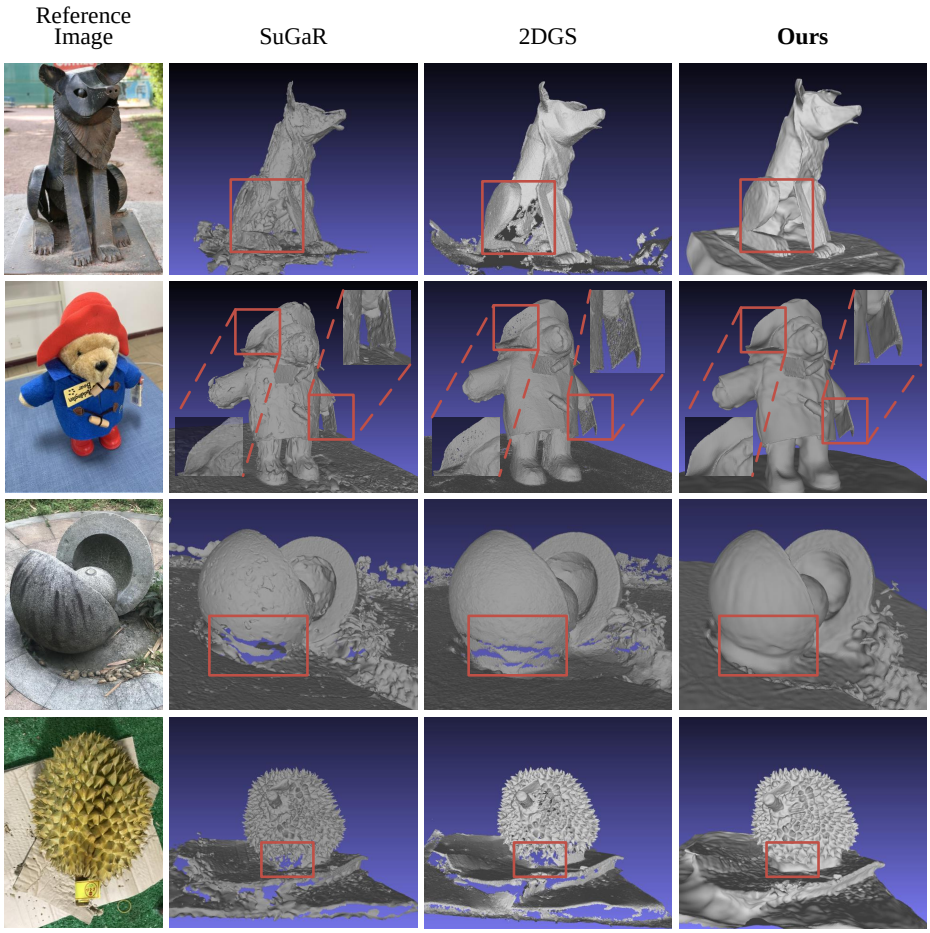

Fig. 4: Qualitative comparison with Gaussian-Splatting-based methods. Unlike the SuGaR [22] and 2DGS [23] methods, our approach facilitates the creation of more continuous surface representations, particularly for surfaces with sparse textures.

4.5 Ablation Studies

Shared Feature Fileds. This article emphasizes the benefits of information sharing between the color field and the SDF field. To investigate this, we conduct two experiments to isolate the information between them to varying degrees. In the first experiment, we do not provide the SDF decoder's output to the color field decoder (denoted as w/o SDF share). In the second experiment, the encoder's output information is not shared (denoted as w/o feature fields share). The results, as shown in Fig. 5, indicate that reducing the shared information decreases the quality of the reconstruction. Therefore, the shared information between different fields facilitates their joint optimization.

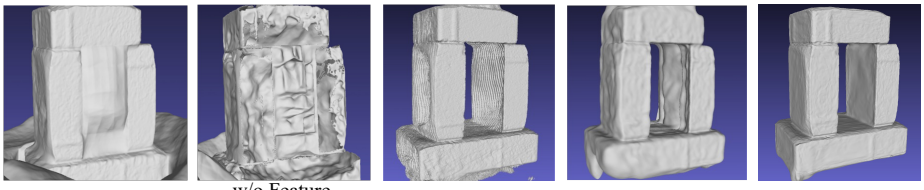

Fig. 5: Ablation Studies. Each component designed is visually important for the final reconstruction result.

The mixed Sinus-ReLU Activation Function. We show other ablation results in Fig. 5 with the classic "Positional Encoding + ReLU" paradigm (denoted as PE+ReLU), in which corrugate artifacts can be found on the reconstructed surface. Then, we use only the sinus function as the activation function for all of the encoders and decoders (denoted as Only Sinus). We find that the reconstruction quality also degrades. Intuitively, we introduce the sinus function to bring some of the frequency encoding ability, while ReLU seems more ideal to represent SDF because its gradient is locally constant and its second derivative is zero. Therefore, we only use the sinus function in the encoder and ReLU in the decoders to compensate for the implicit fields' limited representation ability.

4.6 Limitations

The limitation of our method lies in its inability to reconstruct non-transparent or non-watertight surfaces, an important area for future research that could be addressed by incorporating priors and using representations from multiple implicit fields. Similar to most related methods [44,56,63,64], our approach requires training separate networks for each scene, which is inefficient for practical applications. A potential improvement would be to learn a general implicit field representation that can be fine-tuned for new scenes.

5 Conclusion

In this paper, we propose a new method for 3D multi-view reconstruction from 2D RGB images. Using an Encoder-Decoder paradigm, a shared feature is encoded for the SDF and color field, requiring only a single forward pass for spatial positions, thereby reducing spatial ambiguity. Additionally, the introduction of a mixed Sinus-ReLU activation function helps eliminate corrugated artifacts. Our quantitative and qualitative results demonstrate that this method effectively reconstructs objects with complex topology, especially in textureless and rarely visible regions, while producing high-quality novel views.

Acknowledgments This work is supported by the National Natural Science Foundation of China under Grant U20B2042 and 62076019.

References

1. Agrawal, M., Davis, L.S.: A probabilistic framework for surface reconstruction from multiple images. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. (CVPR) (2001)
2. Aliev, K.A., Sevastopolsky, A., Kolos, M., Ulyanov, D., Lempitsky, V.: Neural point-based graphics. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16. pp. 696–712. Springer (2020)
3. Atzmon, M., Haim, N., Yariv, L., Israelov, O., Maron, H., Lipman, Y.: Controlling neural level sets. Advances in Neural Information Processing Systems(NIPS) (2019)
4. Atzmon, M., Lipman, Y.: Sal: Sign agnostic learning of shapes from raw data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR) (2020)
5. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: Patchmatch: A randomized correspondence algorithm for structural image editing. ACM Trans. Graph. **28**(3), 24 (2009)
6. Broadhurst, A., Drummond, T.W., Cipolla, R.: A probabilistic framework for space carving. In: Proceedings eighth IEEE International Conference on Computer Vision (ICCV) (2001)
7. Campbell, N.D., Vogiatzis, G., Hernández, C., Cipolla, R.: Using multiple hypotheses to improve depth-maps for multi-view stereo. In: Proceedings of the European Conference on Computer Vision(ECCV) (2008)
8. Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H.: Tensorf: Tensorial radiance fields. In: European conference on computer vision. pp. 333–350. Springer (2022)
9. Chen, H., Li, C., Lee, G.H.: Neusg: Neural implicit surface reconstruction with 3d gaussian splatting guidance. arXiv preprint arXiv:2312.00846 (2023)
10. Chen, Z., Zhang, H.: Learning implicit fields for generative shape modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR) (2019)
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(SIGGRAPH) (1996)
12. De Bonet, J.S., Viola, P.: Poxels: Probabilistic voxelized volume reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (1999)
13. Fang, S., Wang, Y., Tsai, Y.H., Yang, Y., Ding, W., Zhou, S., Yang, M.H.: Chat-edit-3d: Interactive 3d scene editing via text prompts. arXiv preprint arXiv:2407.06842 (2024)
14. Fang, S., Wang, Y., Yang, Y., Tsai, Y.H., Ding, W., Zhou, S., Yang, M.H.: Editing 3d scenes via text prompts without retraining. arXiv e-prints pp. arXiv–2309 (2023)
15. Fang, S., Wang, Y., Yang, Y., Xu, W., Wang, H., Ding, W., Zhou, S.: Pvd-al: Progressive volume distillation with active learning for efficient conversion between different nerf architectures. arXiv preprint arXiv:2304.04012 (2023)
16. Fang, S., Xu, W., Wang, H., Yang, Y., Wang, Y., Zhou, S.: One is all: Bridging the gap between neural radiance fields architectures with progressive volume distillation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 597–605 (2023)
17. Fridovich-Keil, S., Yu, A., Tancik, M., Chen, Q., Recht, B., Kanazawa, A.: Plenoxels: Radiance fields without neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5501–5510 (2022)

18. Furukawa, Y., Ponce, J.: Accurate, dense, and robust multiview stereopsis. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2009)
19. Galliani, S., Lasinger, K., Schindler, K.: Gipuma: Massively parallel multi-view stereo reconstruction. Publikationen der Deutschen Gesellschaft für Photogrammetrie, Fernerkundung und Geoinformation e. V **25**(361-369), 2 (2016)
20. Gao, J., Gu, C., Lin, Y., Zhu, H., Cao, X., Zhang, L., Yao, Y.: Relightable 3d gaussian: Real-time point cloud relighting with brdf decomposition and ray tracing. arXiv preprint arXiv:2311.16043 (2023)
21. Gropp, A., Yariv, L., Haim, N., Atzmon, M., Lipman, Y.: Implicit geometric regularization for learning shapes. arXiv preprint arXiv:2002.10099 (2020)
22. Guédon, A., Lepetit, V.: Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5354–5363 (2024)
23. Huang, B., Yu, Z., Chen, A., Geiger, A., Gao, S.: 2d gaussian splatting for geometrically accurate radiance fields. arXiv preprint arXiv:2403.17888 (2024)
24. Huang, P.H., Matzen, K., Kopf, J., Ahuja, N., Huang, J.B.: Deepmvs: Learning multi-view stereopsis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR) (2018)
25. Jensen, R., Dahl, A., Vogiatzis, G., Tola, E., Aanæs, H.: Large scale multi-view stereopsis evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR) (2014)
26. Jiang, Y., Tu, J., Liu, Y., Gao, X., Long, X., Wang, W., Ma, Y.: Gaussianshader: 3d gaussian splatting with shading functions for reflective surfaces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5322–5332 (2024)
27. Kazhdan, M., Bolitho, M., Hoppe, H.: Poisson surface reconstruction. In: Proceedings of the fourth Eurographics Symposium on Geometry Processing(SGP) (2006)
28. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4), 1–14 (2023)
29. Kutulakos, K.N., Seitz, S.M.: A theory of shape by space carving. International Journal of Computer Vision(IJCV) (2000)
30. Leroy, V., Franco, J.S., Boyer, E.: Shape reconstruction using volume sweeping and learned photoconsistency. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
31. Marr, D., Poggio, T.: Cooperative computation of stereo disparity: A cooperative algorithm is derived for extracting disparity information from stereo image pairs. Science **194**(4262), 283–287 (1976)
32. Martin-Brualla, R., Radwan, N., Sajjadi, M.S.M., Barron, J.T., Dosovitskiy, A., Duckworth, D.: NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In: CVPR (2021)
33. Matsuki, H., Murai, R., Kelly, P.H., Davison, A.J.: Gaussian splatting slam. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18039–18048 (2024)
34. Merrell, P., Akbarzadeh, A., Wang, L., Mordohai, P., Frahm, J.M., Yang, R., Nistér, D., Pollefeys, M.: Real-time visibility-based fusion of depth maps. In: Proceedings 11th IEEE International Conference on Computer Vision(ICCV) (2007)
35. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR) (2019)
36. Michalkiewicz, M., Pontes, J.K., Jack, D., Baktashmotlagh, M., Eriksson, A.: Implicit surface representations as layers in neural networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV) (2019)
37. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: Proceedings of the European Conference on Computer Vision(ECCV) (2020)
38. Miller, B., Chen, H., Lai, A., Gkioulekas, I.: Objects as volumes: A stochastic geometry view of opaque solids. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 87–97 (2024)
39. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Trans. Graph. **41**(4), 102:1–102:15 (2022)
40. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: International Conference on Machine Learning(ICML) (2010)
41. Niemeyer, M., Mescheder, L., Oechsle, M., Geiger, A.: Occupancy flow: 4d reconstruction by learning particle dynamics. In: Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV) (2019)
42. Niemeyer, M., Mescheder, L., Oechsle, M., Geiger, A.: Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR) (2020)
43. Oechsle, M., Mescheder, L., Niemeyer, M., Strauss, T., Geiger, A.: Texture fields: Learning texture representations in function space. In: Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV) (2019)
44. Oechsle, M., Peng, S., Geiger, A.: Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV) (2021)
45. 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(CVPR) (2019)
46. Paschalidou, D., Ulusoy, O., Schmitt, C., Van Gool, L., Geiger, A.: Raynet: Learning volumetric 3d reconstruction with ray potentials. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR) (2018)
47. Peng, S., Niemeyer, M., Mescheder, L., Pollefeys, M., Geiger, A.: Convolutional occupancy networks. In: Proceedings of the European Conference on Computer Vision(ECCV) (2020)
48. Picard, Q., Chevobbe, S., Darouich, M., Didier, J.Y.: A survey on real-time 3d scene reconstruction with slam methods in embedded systems. arXiv preprint arXiv:2309.05349 (2023)
49. Riegler, G., Ulusoy, A.O., Bischof, H., Geiger, A.: Octnetfusion: Learning depth fusion from data. In: 2017 International Conference on 3D Vision (3DV) (2017)
50. Samavati, T., Soryani, M.: Deep learning-based 3d reconstruction: a survey. Artificial Intelligence Review **56**(9), 9175–9219 (2023)
51. Schönberger, J.L., Zheng, E., Frahm, J.M., Pollefeys, M.: Pixelwise view selection for unstructured multi-view stereo. In: Proceedings of the European Conference on Computer Vision(ECCV) (2016)
52. Seitz, S.M., Dyer, C.R.: Photorealistic scene reconstruction by voxel coloring. International Journal of Computer Vision(IJCV) (1999)

53. Sitzmann, V., Martel, J., Bergman, A., Lindell, D., Wetzstein, G.: Implicit neural representations with periodic activation functions. Advances in Neural Information Processing Systems(NIPS) (2020)
54. Tola, E., Strecha, C., Fua, P.: Efficient large-scale multi-view stereo for ultra high-resolution image sets. Machine Vision and Applications **23**(5), 903–920 (2012)
55. Vora, A., Gadi Patil, A., Zhang, H.: Divinet: 3d reconstruction from disparate views using neural template regularization. Advances in Neural Information Processing Systems **36** (2024)
56. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. arXiv preprint arXiv:2106.10689 (2021)
57. Wang, Y., Fang, S., Zhang, H., Li, H., Zhang, Z., Zeng, X., Ding, W.: Uav-enerf: Text-driven uav scene editing with neural radiance fields. IEEE Transactions on Geoscience and Remote Sensing (2024)
58. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20310–20320 (2024)
59. Yan, Z., Low, W.F., Chen, Y., Lee, G.H.: Multi-scale 3d gaussian splatting for anti-aliased rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20923–20931 (2024)
60. Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: Mvsnet: Depth inference for unstructured multi-view stereo. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
61. Yao, Y., Luo, Z., Li, S., Shen, T., Fang, T., Quan, L.: Recurrent mvsnet for high-resolution multi-view stereo depth inference. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR) (2019)
62. Yao, Y., Luo, Z., Li, S., Zhang, J., Ren, Y., Zhou, L., Fang, T., Quan, L.: Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1790–1799 (2020)
63. Yariv, L., Gu, J., Kasten, Y., Lipman, Y.: Volume rendering of neural implicit surfaces. Advances in Neural Information Processing Systems(NIPS) (2021)
64. 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(NIPS) (2020)
65. Zhang, K., Riegler, G., Snavely, N., Koltun, V.: Nerf++: Analyzing and improving neural radiance fields. arXiv preprint arXiv:2010.07492 (2020)