Spacetime Surface Regularization for Neural Dynamic Scene Reconstruction

Jaesung Choe[1,2] Christopher Choy[1] Jaesik Park[3] In So Kweon[2] Anima Anandkumar[1,4]

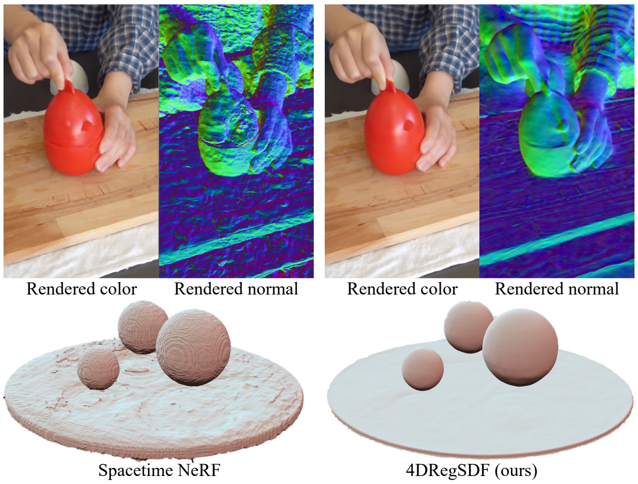
[1]NVIDIA [2]KAIST [3]Seoul National University [4]Caltech

Abstract

We propose an algorithm, 4DRegSDF, for the spacetime surface regularization to improve the fidelity of neural rendering and reconstruction in dynamic scenes. The key idea is to impose local rigidity on the deformable Signed Distance Function (SDF) for temporal coherency. Our approach works by (1) sampling points on the deformed surface by taking gradient steps toward the steepest direction along SDF, (2) extracting differential surface geometry, such as tangent plane or curvature, at each sample, and (3) adjusting the local rigidity at different timestamps. This enables our dynamic surface regularization to align 4D spacetime geometry via 3D canonical space more accurately. Experiments demonstrate that our 4DRegSDF achieves state-of-the-art performance in both reconstruction and rendering quality over synthetic and real-world datasets. https://4dregsdf.github.io/

1. Introduction

Reconstructing 4D dynamic scenes from natural videos is one of the fundamental problems of vision tasks that are widely used in various fields, such as robotics, autonomous driving, virtual reality, and many more [6, 8, 14, 24, 32]. It requires an understanding of not just the geometry of scenes but also the texture, illumination, material properties, and refraction for photo-realistic rendering. However, because of the difficulty in reconstructing static 3D geometry alone, learning the dynamics jointly has been a challenge that most conventional methods decomposed the task into multiple stages of geometry reconstruction, motion representation, texture mapping, and illumination estimation [6, 15, 33, 49, 68]. Such sequential pipelines that first reconstruct geometry and learn motion separately are not robust to many failure modes of each stage, resulting in inaccurate reconstruction, which is one of the major bottlenecks for photo-realistic and geometry-oriented spacetime scene understanding.


Rendered color Rendered normal Rendered color Rendered normal

Spacetime NeRF 4DRegSDF (ours)

Figure 1. We present an implicit surface regularization method that fully utilizes the properties of SDF for robust shape reconstruction and novel-view rendering in dynamic scenes.

The recent advances in neural implicit representations, however, made accurate 3D reconstruction with differentiable volumetric rendering possible [40,41,46,50]. Mildenhall *et al*. [41] use sinusoidal encoding on the position as an input to multi-layered perceptrons (MLPs) to learn appearance and geometry of the target static scene. Specifically for photo-realistic rendering, neural radiance fields [4, 41, 54, 61, 63, 71, 79] mostly use density and color as the main output. However, the density representation can result in artifacts such as floaters and blurry boundaries since the density representation has high degrees of freedom [67, 70]. For accurate 3D reconstruction, the Signed Distance Function (SDF) started to gain traction for its high fidelity 3D surface reconstruction [26, 47, 70, 76, 81]. One of the key component for SDF learning is the Eikonol loss [25] which regularizes the SDF to have valid gradient. This acts as a strong regularization not just on the surface but on the entirety of the empty space.

However, the 4D surface reconstruction from monocular RGB videos – which we reconstruct a dynamic scene – is still a highly ill-posed problem where we have to recon-

Work done during the internship at NVIDIA.

struct geometry and the dynamics of the scene simultaneously. Recent studies proposed to decompose the 4D space into a static geometry/texture (canonical model) and deformation to embed physical constraints into the 4D representation [51, 54]. The decomposition of the geometry could regularize the 4D space, but many papers use additional cues such as flow, or depth prediction [16, 21, 31, 34, 35, 69, 72, 78]. Still, the reconstruction is an ill-posed problem and it often fails for two reasons: (1) learning deformation is sensitive to topological variation and illumination change that often requires energy minimization terms for stable training. [28, 51, 60, 66]; (2) using the density representation – which has high degrees of freedom – could result in overfitting, floaters, and blurry reconstructions [29, 37, 44] as shown in Fig. 1. Thus, regularizing the spacetime domain is one of the most crucial parts of the 4D scene understanding in terms of geometry and graphics.

In this paper, we propose an algorithm for spacetime surface regularization that improves the fidelity of neural rendering and reconstruction in dynamic scenes. The key idea is to enforce the local rigidity on the deformable Signed Distance Function (SDF) for the 4D geometry representation. This allows us to (1) regularize the reconstruction using more explicit surface representation instead of fuzzy density measure [41]; (2) sample geometry efficiently for surface regularizations. Specifically, previous studies [37, 47, 70] show that Signed Distance Function can reconstruct geometries accurately with fewer images since learning the distance to the nearest surface to represent a surface is strong regularizations compared to the density function. In addition, the signed distance encodes not just distance but direction to the closest surface from gradients, thus allowing efficient surface sampling. Then, the sampled points are used to minimize the deformation energy and align 4D spacetime geometry via 3D canonical space.

We use three datasets to evaluate our method: one synthetic dataset from Pumarola *et al*. [54], and two real datasets: Park *et al*. [52] and Gao *et al*. [22]. Our method achieves state-of-the-art performance on both geometry representation and appearance as visualized in Fig. 1. We summarize our contributions as follows:

- Extends the energy minimization in conventional 4D surface reconstruction studies into learning neural dynamic scene reconstruction.

- Regularize geometric properties of the 4D surface by the total variation of the curvature (TVC) and the absolute curvature of the SDF.

- Robust performance of rendering and reconstruction under the dynamic few-shot setup compared to recent implicit geometric regularization studies [20, 44].

2. Related works

Conventional reconstruction and rendering. Video often contains time-varying properties, such as illumination change, geometry deformation, topology changes, and rigid or deformable motion. Such time-varying elements result in multi-view inconsistency among different frames and increase the level of difficulties of the video rendering task. For this issue, previous studies often require exhaustive multi-camera setup [5, 13, 48, 82] or synchronized multiple videos [2, 3]. None-learning based methods [1, 28, 60, 62] commonly assume the coarse geometry from RGB-D priors to introduce the energy regularization for the realistic geometry and photometric description of the target dynamic scenes. For instance, terzopoulos *et al*. [66] use elastic energy to deform the surface, and Slavcheva *et al*. [60] propose three typical energy minimization to prevent uncontrolled deformations. These ideas are related to the local rigidity to reconstruct plausible and geometrically-meaningful surface [1, 62] along the spatio-temporal domain.

Neural implicit reconstruction and rendering. Recently, neural scene representation [9, 23, 40, 43, 45, 50, 53, 64] gets noticeable attention with their promising quality of implicit shape representations, such as occupancy or signed distance function. Not just geometry, the studies [36, 46, 58, 59] handle both geometry and appearance from multi-view images by neural implicit representation. In terms of the novel-view synthesis task, Mildenhall *et al*. [41] demonstrates the impressive quality of rendered images. This paper takes a 5D radiance field (3D location and 2D viewing direction) to infer color and density. Then, it accumulates per-point inferences along each ray to determine pixel color. Such formulation makes it possible to train MLPs for learning geometry and appearance simultaneously. Based on this design, the series of following papers address the remaining problems of the pioneering paper [41], such as generalization [7, 57, 71], few-shot images [29, 44, 80], or surface reconstruction by SDF [70, 76, 77]. Nonetheless, these methods mostly focus on targeting static scenes without consideration of the moving objects or temporal change.

Neural video rendering. More recently, several spacetime view synthesis studies [21, 34, 51, 51, 54, 55, 72] adopt neural implicit representations to render novel view synthesis from a video. Typically, these studies can be divided into two dominant strategies: deformable NeRFs [19, 35, 51, 54] and prior-based video NeRFs [21, 34, 72]. Concurrent works [19, 35] propose fast deformable neural rendering methods by using voxel representations. Typically without relying on off-the-shelf information, such as monocular depth maps [34, 78], surface normal maps [81], or optical flow [35], we aim to render and reconstruct scenes purely from posed images.

Regularization for neural implicit representation. Relying on the nature of the geometry, a series of papers [29, 44, 55] present the effectiveness of the regularization in learning neural fields. Under the few-shot setup, Niemeyer *et al.* [44] regularize the unobserved viewpoints via normalizing flow models. Kim *et al.* [29] apply entropy loss to alleviate reconstruction inconsistency. Fu *et al.* [20] apply multi-view consistency for the static surface reconstruction. In light of the growing popularity of spatio-temporal surface reconstruction and neural rendering, this paper proposes a novel approach to regulating the spacetime surface through the use of 3D canonical space. We have observed that the direct application of the SDF to this deformation field can result in low-fidelity reconstruction and rendering due to ill-posed problems in dynamic scenes. To overcome this issue, we highlight the importance of our regularization to achieve precise reconstruction and rendering.

3. Preliminary

Neural radiance field [41] allows fully differentiable learning of geometry and appearance by learning color and density from RGB images. Recent studies [42, 70] on surface reconstruction utilize Signed Distance Function (SDF) to represent scene geometry since The SDF representation has specific favorable characteristics that give it an edge for this particular objective.

Differential surface geometry in SDF. The Signed Distance Function (SDF) stands as a predominant representation for implicit geometry. We can derive mathematical attributes on a crisp surface from SDF that could be challenging to derive from density-based representations: surface normal vector $\mathbf{n} = \frac{\partial s}{\partial \mathbf{p}}$, tangent plane projection matrix $P = I - \mathbf{n}^\top \mathbf{n}$, and surface curvature through the Hessian matrix $H = \frac{\partial^2 s}{\partial \mathbf{p}^2}$ [73] Note that I denotes the identity matrix, s the SDF value, and $\mathbf{p}$ a 3D coordinate. We can derive these mathematical attributes because SDF represents the distance to the closest surface at every point. We can force this property for implicit geometry by using the Eikonal loss [25], which forces the norm of gradients of SDF $(\|\mathbf{n}\|_2 - 1)^2$ to be 1, and the loss has been widely used as an important regularization for 3D reconstruction methods [20, 25, 70, 76].

NeuS [70] uses the same volumetric rendering equation in NeRF [41]. However, the opaqueness α is derived from the SDF value s as $\alpha{=}1{-}\exp\big(-\int_{t_{i+1}}^{t_i} \rho(t)dt\big)$ where ρ is the density function $\rho{=}\max\big(\frac{s-\Delta s}{s+\Delta s}, 0\big)$ and Δs is the displacement of the SDF. Following NeRF [41], NeuS utilizes the volumetric rendering as:

$$\hat{C}(\mathbf{r}) = \sum_{i=1}^{N} T_i \alpha_i \mathbf{c}_i \;\; \text{s.t.} \;\; T_i = \exp(-\sum_{j=1}^{i-1} \sigma_i \delta_i), \quad (1)$$

where $\mathbf{c}_i$ denotes the color of the i-th sample point $\mathbf{p}_i$, α_i is the opaqueness over distance δ_i on the ray $\mathbf{r} = \{\mathbf{p}_i\}$. T_i is the cumulative transmittance along the ray. Based on this formulation, NeuS [70] is trained purely from posed images to describe the static scene surface using the SDF.

Deformable NeRF [51, 54] is one way of novel-view synthesis methods in the 4D dynamic scenes. The idea is to decompose the 4D deformed space $[\mathbf{p}; t]$ into the 3D canonical geometry $\mathbf{p_c}$ and the temporal deformation $\Delta\mathbf{p}$ as:

$$\mathbf{p}_c = \text{Deform}([\mathbf{p}; t]) = \mathbf{p} + \Delta\mathbf{p}, \quad (2)$$

where $\Delta\mathbf{p}(=h([\mathbf{p}; t]))$ is the amount of space deformation, a function $h(\cdot)$ represents a deformation field, and the deformation field transports a point $[\mathbf{p}; t]$ to the canonical coordinate $\mathbf{p}_c = [x_c, y_c, z_c]$. Although progress has been made in the field of dynamic neural rendering, we address the importance of the spacetime surface regularization for learning appearance, geometry, and motion in dynamic scenes as visualized in Fig. 1.

4. Method

This section presents our framework, 4DRegSDF. Our network represents the 4D spacetime domain as a combination of the static geometry in Signed Distance Function and the deformation field. Given a set of posed frames from a monocular RGB video, our network is trained to encode a dynamic scene (Sec. 4.1) and apply our spacetime surface regularization (Sec. 4.2). In contrast to previous video neural field papers that use priors such as depths or flows [21, 34, 35], our method rely only on posed image frames from a video as input to the system making the system simple and efficient.

4.1. Deformable Signed Distance Function

Given a point $\mathbf{p}$ at time t within the deformed space and its viewing direction $\mathbf{d}$, our model f predicts the color $\mathbf{c}$, the SDF value s, and the deformation delta $\Delta\mathbf{p}$. This can be expressed as:

$$[s; \Delta\mathbf{p}; \mathbf{c}] = f([\mathbf{p}; t], \mathbf{d}), \quad (3)$$

As visualized in Fig. 3, we use the same volumetric rendering on the SDF (Eq. 1) [70] to get color value per each ray $\mathbf{r}$, and minimize the photometric loss $\mathcal{L}_\text{c}$ as,

$$\mathcal{L}_\text{c} = \Big| C(\mathbf{r}) - \hat{C}(\mathbf{r}) \Big|_1, \quad (4)$$

where C and $\hat{C}$ are ground truth color and predicted color, respectively, along a query ray $\mathbf{r}$.

4.2. Spacetime surface regularization

Reconstructing dynamic surfaces from videos presents challenges due to its inherent complexity, often leading to

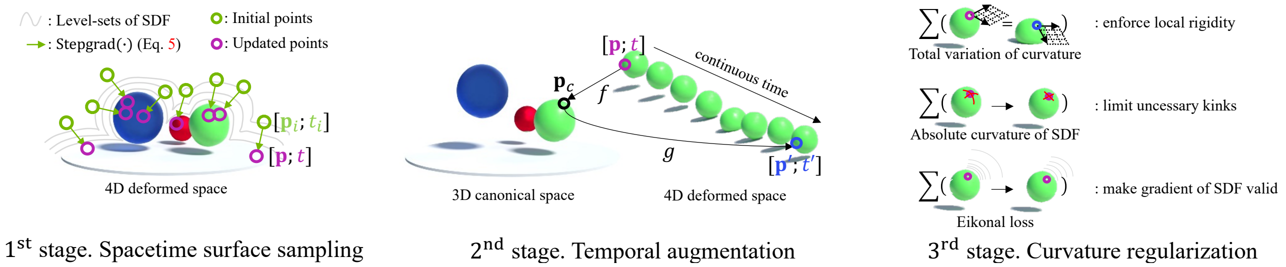


Figure 2. Surface sampling for our spacetime surface regularization in three stages.

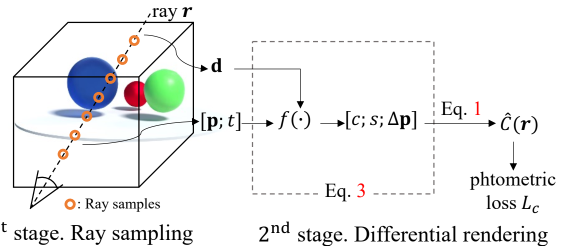


Figure 3. Ray sampling for neural rendering [70].

visual artifacts as shown in Fig. 6. Accordingly, designing a robust regularization is one of the critical problems for successful 4D reconstruction [1, 51, 60, 62, 66, 73].

In this section, we adopt three regularizations to enhance the learning of deformable Signed Distance Function from videos: 1. total variation of curvature, 2. the absolute curvature space-time, and 3. the Eikonal loss [25]. The first regularization aims to limit the amount of change in curvature induced by motion and the second regularization flattens out the 4D reconstruction and limits unnecessary kinks. The last regularization [25] enforces the gradient of SDF to be valid. First two regularizations are related to the properties of surface and can be computed only on the surface. We elaborate on the sampling strategy and how to augment the samples in continuous space and time, and lastly present the regularizations. Overall pipeline is visualized in Fig. 2.

Spacetime surface sampling. The most commonly used sampling in neural fields is ray casting where you select points along a ray for rendering. Many regularizations proposed for NeRF also use samples along a ray [20, 29, 37, 44]. However, surface constraints we adopt require samples on the surface of objects no along rays. For this, we use the definition of the SDF for surface sampling.

The SDF is defined as the (signed) distance to the closest surface from the query point. In static scenes, the surface normal indicates the direction of the closest surface [12, 39, 73]. However, in 4D dynamic scenes, time is another component to determine the closest surface. In our setup, our network $f(\cdot)$ is trained to infer the SDF value s from a query $[\mathbf{p}; t]$. Accordingly, the gradient step towards the *closest* surface naturally involves and spatial update as well as the time update. Given randomly initialized point p_i at the time t_i, each query point travels along the spacetime SDF 'hills' to be located at the isosurface as:

Algorithm 1 Spacetime surface regularization

```
Input: Randomly initialized 4D point $[\mathbf{p}_i; t_i]$.
Output: Total regularization $\mathcal{L}_e$.
 1: function 4DRegSDF($[\mathbf{p}_i, t_i]$)
 2:     $[\mathbf{p}, t] \leftarrow [\mathbf{p}_i, t_i]$
 3:     for iterations do
 4:         $[\mathbf{p}, t] \leftarrow$ StepGrad($[\mathbf{p}; t]$)          ▷ Eq. 5
 5:     end for
 6:     $\mathbf{p}_c \leftarrow$ Deform($[\mathbf{p}; t]$)                  ▷ Eq. 2
 7:     $t' \leftarrow$ Random(0, 1)                       ▷ Time augmentation
 8:     $\mathbf{p}' \leftarrow$ InvDeform($[\mathbf{p}_c; t']$)             ▷ Eq. 6
 9:     $\mathcal{L}_e \leftarrow$ ComputeLoss($[\mathbf{p}; t], [\mathbf{p}'; t']$)   ▷ Eq. 9
10:     return $\mathcal{L}_e$
11: end function
```

$$[\mathbf{p}; t] = \text{StepGrad}([\mathbf{p}_i; t_i]) = [\mathbf{p}_i; t_i] - s \cdot \frac{\partial s}{\partial [\mathbf{p}_i; t_i]}, \quad (5)$$

where $\frac{\partial s}{\partial [\mathbf{p}; t]}$ is the gradient of the SDF value s in terms of $[\mathbf{p}; t]$. By doing so, we can regularize the 4D surface in a wide-spectrum time domain including unobserved timestamps within training samples. Thus, by iteratively conducting this operation above, we move the initial 4D point $[\mathbf{p}_i; t_i]$ to $[\mathbf{p}; t]$ closer to the isosurface having smaller SDF value, as described in L. 2-5 of Alg. 1.

Temporal augmentation. Given the surface samples $[\mathbf{p}; t]$ from Eq. 5, we enforce the local rigidity to the surface samples for temporal coherency. First, we augment the time component t into randomly selected t' as visualized in Fig. 2-(b). Specifically, we use the inverse-deformation network g [35, 45, 74] which is formulated as:

$$\mathbf{p}' = \text{InvDeform}([\mathbf{p}_c; t']) = g(\mathbf{p}_c, t'), \quad (6)$$

where $\mathbf{p}_c$ indicates the points at the canonical space from $[\mathbf{p}; t]$ (Eq. 3). $\mathbf{p}'$ is a inverse-deformed point from $\mathbf{p}_c$ at the augmented time t' as stated in L. 6-8 of Alg. 1. Through our temporal augmentation stage, we get a pair of two surface samples that correspond to each other, $([\mathbf{p}; t], [\mathbf{p}'; t'])$. This pair will be used for evaluating spatio-temporal constraints the next section.

Regularization on the spacetime curvature. Inspired by the conventional 4D reconstruction studies [1,60,62,66], we regularize geometric properties of dynamic surfaces: the *total variation of curvature* (TVC) and the absolute curvature of SDF surface on continuous time. The TVC measures the change of the curvature at different time steps and by minimizing the change, we assumes that surface should deform as little as possible while satisfying video observations. On the contrary, the absolute curvature measures curvature values and thus regularizes the complexity of the reconstruction and movement by minimizing the distortion in space. For the TVC, we adopt the stretching loss [73] where it minimizes the difference of inner products of two tangent vectors at two different time steps: v_1, v_2 at t and v_1', v_2' at t'. The tangent vector is a remainder of a vector on surface $v = P\mathbf{p}$ where $P(= I - \mathbf{n}\mathbf{n}^\top)$ is the tangential projection matrix and $\mathbf{n}(= \frac{\partial s}{\partial \mathbf{p}})$ is the surface normal vector at a point $\mathbf{p}$. Let J_1, J_2 be the Jacobians transforming v_1, v_2 to v_1', v_2' respectively. Then, we compute the TVC as:

$$|v_1^\top v_2 - v_1'^\top v_2| = |\mathbf{p}_1^\top P_1^\top P_2 \mathbf{p}_2 - \mathbf{p}_1^\top P_1^\top J_1^\top J_2 P_2 \mathbf{p}_2|, \tag{7}$$

$$\propto |P_1^T (I - J_1^T J_2) P_2|, \tag{8}$$

The absolute curvature of the 4D SDF is the L1 norm of the Hessian matrix $H = \frac{\partial^2 s}{\partial \mathbf{p}^2}$. By minimizing the absolute curvature, the reconstruction removes unnecessary curvatures or kinks in the reconstructions making the surface as smooth as possible. The total regularization is:

$$\mathcal{L}_e(\mathbf{p}, \hat{\mathbf{p}}) = \|P^\top (I - J_g J_f) P\|_2 + \|H\|_1, \tag{9}$$

where J_g and J_f are the Jacobians that transformed points via the deformation network $f(\cdot)$ and the inverse deformation network $g(\cdot)$, respectively.

We summarize our spacetime surface regularization in Algorithm 1. We achieve high-quality reconstruction results that are geometrically simple while they match video observations through neural rendering and minimizing the photometric loss. Finally, we train our network by combining the proposed losses as below,

$$\mathcal{L} = \lambda_c \mathcal{L}_c + \lambda_{cc} \mathcal{L}_{cc} + \lambda_{\text{e}} \mathcal{L}_{\text{e}} + \lambda_{\text{eik}} \mathcal{L}_{\text{eik}}. \tag{10}$$

where $\lambda_c = 1$, $\lambda_{\text{cc}} = 0.01$, $\lambda_e = 0.001$, and $\lambda_{\text{eik}} = 0.01$. $\mathcal{L}_{\text{eik}}$ represents the Eikonal loss [25] following NeuS [70]. Also, we enforce cycle consistency loss in learning forward/inverse deformation fields by $\mathcal{L}_{cc} = \|\mathbf{p} - g(\mathbf{p}_c, t)\|_1$ [35,45,74].

5. Experiments

5.1. Novel-view synthesis

Given a set of posed images, the novel-view synthesis aims to render images from viewpoints that are not in the part of the provided posed images. Since our 4DRegSDF also follows the neural rendering schemes, we compare the quality of rendered images with the recent spacetime NeRF papers on novel-view synthesis evaluation.

Dataset. We utilize the Dynamic NeRF Synthetic dataset provided by D-NeRF [54] and the real-world scenes from HyperNeRF dataset [52]. These datasets offer natural monocular videos with dynamic frames and continuous camera motion. We follow the official train/test split for training and evaluation over ours and related studies. Following the conventions [41,51,54], we evaluate the rendering performance through PSNR, SSIM, and LPIPS.

Comparison. We compare our method with a few state-of-the-art 4D NeRF papers, D-NeRF [54], TiNeuVox [19], which also do not rely on the prior information [21,34,35]. Also, we extend NeuS [70] for the 4D dynamic reconstruction task by adopting a deformation network from D-NeRF [54]. We denote this deformable NeuS as 'NeuS+D' and use it One of the concurrent works, DeVRF [35], also proposes a deformation-based dynamic neural rendering task. However, the paper requires additional multi-view static scenes for initialization so we did not include it in our experiment.

When compared to existing research, our method achieves state-of-the-art performance among the recent spacetime NeRF methods as shown in Table 1 and Table 2. Many of these studies rely on view-dependent surface representations and the density function [41]. In contrast, our method utilizes the SDF representation, allowing us to more accurately represent geometry. This distinction influences the rendering quality. Notably, in terms of the qualitative reconstruction results, the previous state-of-the-art method struggles to accurately capture surface details as illustrated in Fig. 4 and Fig. 5. We suspect this limitation is rooted in the inherent issues of the density representation. As a result, many of the earlier spacetime NeRF studies place a heavy emphasis on the rendering quality. However, we claim that by imposing scene geometry constraints through our spacetime surface regularization, we can further improve the quality of the geometry under the precise geometry presentation, namely SDF.

5.2. Scene reconstruction

Dataset. Dycheck dataset [22] is one of the few datasets that provide ground truth geometry information for dynamic scenes. Yet, the ground truth depth maps are only provided for the training data split and we could not evalu-

	Dynamic Synthetic NeRF dataset [54]															
	Bouncing balls		Hell warrior		Hook		Jumping jacks		Lego		Mutant		Stand up		T-rex	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
NeRF [41]	20.26	0.91	13.51	0.81	16.65	0.84	18.28	0.88	20.30	0.79	20.31	0.91	18.19	0.89	24.49	0.93
D-NeRF [54]	38.93	0.98	25.02	0.95	29.25	0.96	32.80	0.98	21.64	0.83	31.29	0.96	32.79	0.98	31.75	0.97
TiNeuVox [19]	40.73	**0.99**	**28.17**	**0.97**	31.45	0.97	**34.23**	0.98	25.02	0.92	33.61	0.98	**35.43**	**0.99**	32.70	0.98
TorchNGP [65]	40.44	**0.99**	24.54	0.94	31.46	0.97	31.45	0.95	25.07	0.93	**35.86**	**0.99**	35.29	**0.99**	31.35	0.97
NeuS+D [70]	36.98	0.96	23.77	0.94	28.09	0.94	28.28	0.93	24.17	0.90	30.17	0.94	30.99	0.95	24.94	0.93
Ours	**40.89**	**0.99**	25.72	0.95	**32.20**	**0.98**	33.63	**0.99**	**25.10**	**0.94**	31.83	0.96	33.60	**0.99**	**33.64**	**0.99**

Table 1. Rendering evaluation in the D-NeRF synthetic dataset. We use PSNR (↑) and SSIM (↑) for the metric.

	HyperNeRF dataset [52]									
	Broom		3D Printer		Chicken		Peel-banana		MEAN	
	PSNR	MS-SSIM	PSNR	MS-SSIM	PSNR	MS-SSIM	PSNR	MS-SSIM	PSNR	MS-SSIM
NeRF [41]	19.9	0.653	20.7	0.780	19.9	0.777	20.0	0.769	20.1	0.745
NV [36]	17.7	0.623	16.2	0.665	17.6	0.615	15.9	0.380	16.9	0.571
Nerfies [51]	19.2	0.567	20.6	0.830	26.7	0.943	22.4	0.872	22.2	0.803
TorchNGP [65]	19.8	0.571	21.1	0.841	27.6	0.946	21.1	0.851	23.5	0.815
TiNeuVox [19]	21.5	0.686	22.8	0.841	28.3	0.947	24.4	0.873	24.3	0.837
NeuS+D [70]	21.9	0.689	22.1	0.839	25.9	0.933	24.6	0.875	23.6	0.834
Ours	**23.3**	**0.702**	**23.9**	**0.859**	**29.8**	**0.965**	**25.4**	**0.881**	**25.3**	**0.851**

Table 2. Quantitative evaluation in the HyperNeRF dataset [52].

	Dycheck dataset [22]															
	Creeper				Haru-sit				Paper-windmill				Sriracha-tree			
	RMSE↓	CD↓	F1↑	PSNR↑	RMSE↓	CD↓	F1↑	PSNR↑	RMSE↓	CD↓	F1↑	PSNR↑	RMSE↓	CD↓	F1↑	PSNR↑
TorchNGP [65]	0.340	0.050	87.78	17.32	0.502	0.081	76.33	27.59	0.311	0.057	81.98	23.3	0.311	0.052	81.44	27.9
TiNeuVox [19]	0.345	**0.048**	88.42	16.69	0.634	0.080	75.18	26.11	0.313	0.053	82.29	**24.96**	0.325	0.057	81.50	**30.57**
NeuS+D [70]	0.350	0.071	74.18	17.05	**0.491**	0.118	60.13	27.88	**0.295**	0.038	91.32	22.52	**0.283**	0.045	85.23	26.83
Ours	**0.323**	0.064	82.36	**17.57**	0.494	**0.077**	**78.53**	**28.29**	0.310	**0.051**	**83.24**	24.26	0.290	**0.042**	**92.25**	30.66

Table 3. Evaluation in the Dycheck dataset [22] Note that CD (↓) means Chamfer Distance and F1 (↑) is F1-score.

ate geometry reconstruction quality on the test split. Accordingly, we subsample the official train data to form a new train/test split. The new split consists of 75% training data and 25% test data per scene. For evaluation metrics, we follow the three geometric measures, RMSE, Chamfer Distance, and F1-score. RMSE is a widely-adopted metric for depthmap prediction evaluation [10, 17, 27, 56]. Chamfer distance [11, 18] measures the distance between the ground truth point cloud and the prediction. Meanwhile, the F1-score measures the 3D volume occupancy between the ground truth and the predictions [11, 30, 75].

Comparison. To evaluate geometric reconstruction quality, we evaluate on both view specific depth map prediction task and the 3D geometry reconstruction task. For depth map prediction task, we extract the depth map at the zero-crossing of SDF. For density-based NeRFs [19], we render depth maps by computing the expected depth. For the final metric, measure the RMSE in the neural volume rendering schemes [41, 70]. For 3D reconstruction evaluation, we follow the surface extraction pipeline provided by the official D-NeRF implementation for density-based methods [19, 54]. In our case, we extract mesh at the zero-crossings of SDF by using marching-cube algorithm [38] for 3D geometric evaluation. For evaluation metrics, we utilize the Chamfer Distance (CD ↓), F1-score (F1 ↑), and RMSE ↓.

Based on this formulation, we calculate the 2D/3D geometric performance by ours and recent studies [19]. As described in Table 3, our method still consistently outperforms the quality of 4D surface reconstruction tasks. We will state the detailed ablation study in the following section, however, in short, the dominant performance improvements are driven by the plausible surface representation by Signed Distance Function and the proposed surface regularization schemes. Finally, our 4DRegSDF successfully renders and reconstructs the 4D dynamic scenes thanks to the plausible geometry representation and learning schemes

6. Ablation studies

In this section, we provide additional experiments to analyze each module of the 4DRegSDF. Note that additional ablation results are included in the supplementary material.

Effectiveness of surface regularization. Surface regularization is the key to making 4D surfaces plausible shapes while learning deformation. When comparing with Fig. 6-(b) and (c), we found that $\mathcal{L}_e$ affects the structure of the reconstructed table. It implies that our surface regularization takes a role in shaping plausible structures by aligning geometry between surfaces at different times.

Comparison with different regularization schemes. We compare our surface regularization with previous regularization schemes, *e.g*. RegNeRF [44] and GeoNeuS [20], as in Table 5. Based on these results, we claim that our spacetime surface regularization is particularly promising compared to previous regularization strategies. Also, we pro-

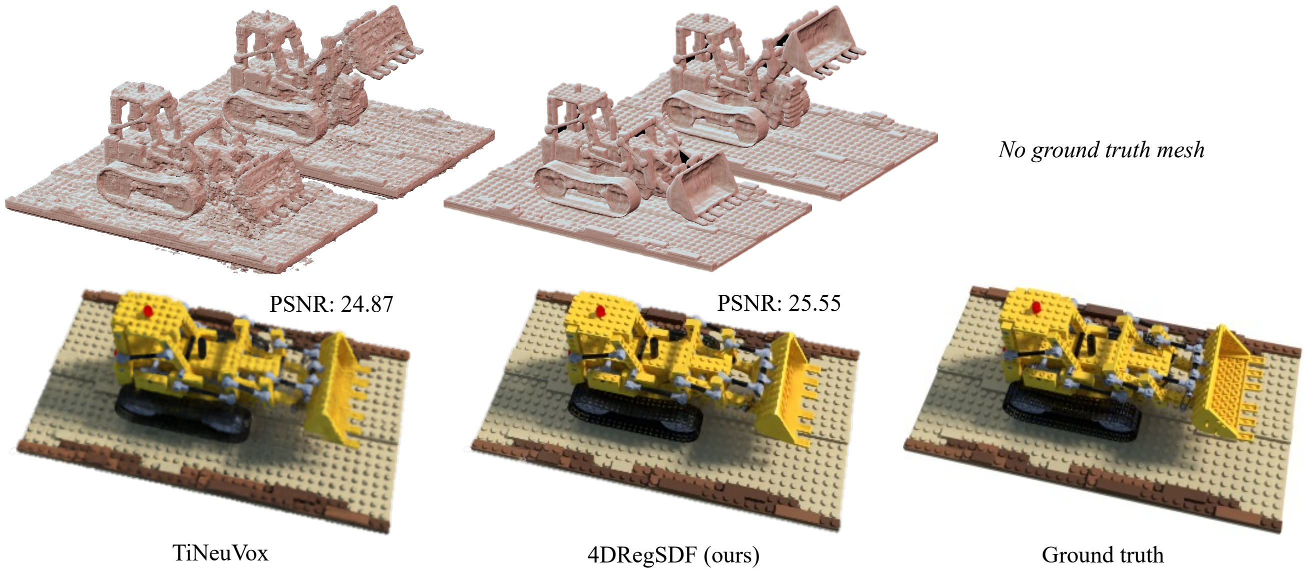


Figure 4. Reconstruction results (mesh) with rendered images from our 4DRegSDF and previous state-of-the-art method, TiNeuVox [19].

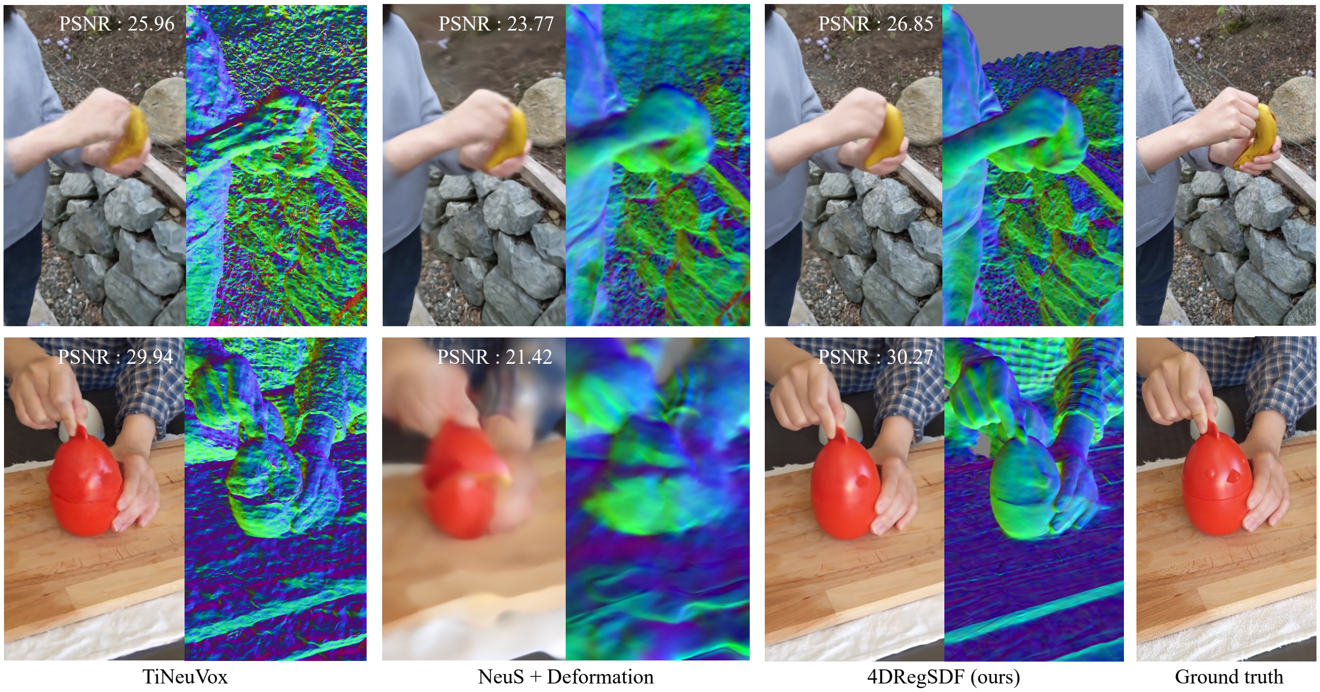
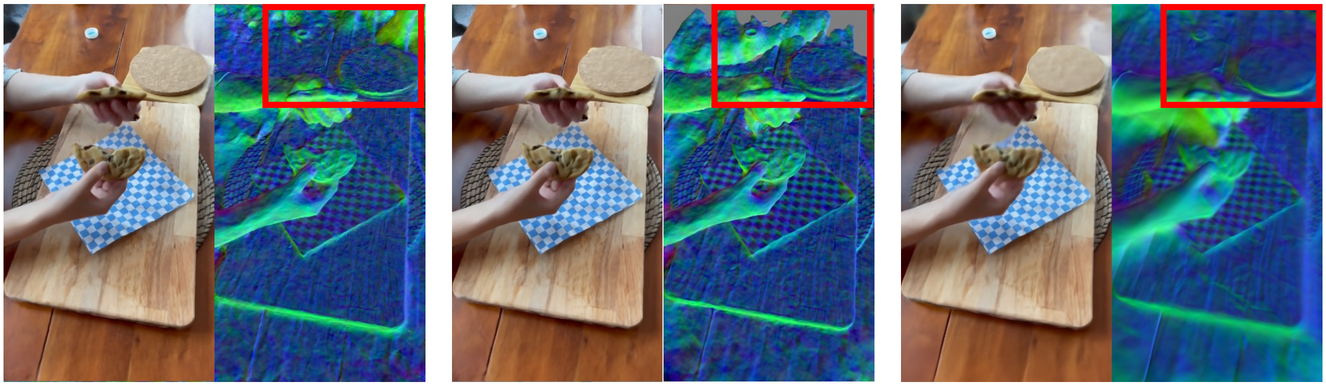


Figure 5. Qualitative results in HyperNeRF dataset [52]. We compare previous state-of-the-art method, TiNeuVox [19], NeuS [70] with deformation [54] and our 4DRegSDF. Left: rendered color images, right: rendered surface normal maps.

vide details of our surface regularization in Table 6. Note that we calculate avg. PSNR from HyperNeRF dataset [52], and avg. RMSE from DyCheck dataset [22].

Deformation network design. Unlike deformable NeRF methods [51, 54] that sequentially process deformation and rendering (Eq. 2), our 4DRegSDF proposes to jointly learn deformation and SDF through a unified network f (Eq. 3) as stated in Sec. 4.1. It turns out that our network design shows the better quality of surface normal maps as visualized in Fig. 6-(a) and (c). Such architecture design can



(a) NeuS + Deformation

(b) 4DRegSDF w/o $\mathcal{L}_e$

(c) 4DRegSDF with $\mathcal{L}_e$ (ours)

Figure 6. Qualitative comparison on (a) NeuS [70] combined with a deformable network [54], (b) ours without surface regularization, and our final 4DRegSDF method. Left: rendered color image, right: rendered surface normal map.

	Ratio of trained data					
	100% (Tables 2,3)			25%		
	PSNR	MS-SSIM	CD	PSNR	MS-SSIM	CD
TiNeuVox [19]	22.8	0.841	0.057	18.1	0.702	0.085
NeuS + D [70]	22.1	0.839	0.045	19.4	0.780	0.064
Ours w/o Reg.	22.4	0.839	0.042	18.1	0.770	0.060
Ours	23.9	0.859	0.042	21.1	0.852	0.058

Table 4. Ablation study on few-shot frames. We choose the two scenes, 3D printer and Sriracha-tree, for the photometric and geometry evaluation. Ratio indicates the percentage of the data that we used for training the methods.

Network	Preserve (✓)			Evaluation	
	RegNeRF	GeoNeuS	Ours	PSNR ↑	RMSE ↓
4DRegSDF	✓			24.1	0.414
		✓		24.7	0.384
			✓	**25.3**	**0.354**

Table 5. Comparison of various regularization methods. RegNeRF [44] regularizes unobserved viewpoints and GeoNeuS [20] applies multi-view consistency as constraint.

be one choice that reduces deformation ambiguity by learning jointly with SDF within a unified network f and having cyclic deformation constraints ($\mathcal{L}_{\text{cc}}$).

Sparse view. To further verify the benefit of our surface regularization schemes, we conduct additional experiments regarding the sparse input views [29, 44]. For experiments, we use '3D printer' and 'Sriracha-tree' for evaluation. As stated in Table 4, when the training data is reduced by 25%, our method shows less performance drop when using a regularization scheme, showing less than 10% performance drop. Such results consistently support the claim that regularization techniques are effective for learning radiance fields under the given sparse views [29, 44].

7. Conclusion

We introduce using geometric spacetime regularizations for neural rendering and surface reconstruction in dynamic scenes from a monocular video. While previous dynamic NeRF methods have primarily targeted rendering quality, our approach places greater emphasis on providing accurate geometry representation via surface regularization. Our method has demonstrated promising performance in both dynamic novel-view synthesis and 4D surface reconstruction tasks. However, a notable constraint of this work is it limited capability in processing extended videos and we will extend this work for long format videos.

Network	Preserve (✓)			Evaluation	
	TVC	Abs	4D sample	PSNR ↑	RMSE ↓
4DRegSDF	✓		✓	24.4	0.392
		✓	✓	24.9	0.379
	✓	✓	✓	**25.3**	**0.354**

Table 6. Ablation study on regularization. We decompose $\mathcal{L}_e$ (Eq. 9) into total variation of curvature (TVC) and absolute curvature of SDF (ABS). 4D sample indicates Eq. 5.

Acknowledgement. This work was supported by IITP grants (No.2021-0-02068: AI Innovation Hub and No.2023R1A1C200781211) funded by the Korea government (MSIT).

References

[1] Marc Alexa, Daniel Cohen-Or, and David Levin. As-rigid-as-possible shape interpolation. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 157–164, 2000.

[2] Luca Ballan, Gabriel J Brostow, Jens Puwein, and Marc Pollefeys. Unstructured video-based rendering: Interactive exploration of casually captured videos. In *ACM SIGGRAPH 2010 papers*, pages 1–11. 2010.

[3] Aayush Bansal, Minh Vo, Yaser Sheikh, Deva Ramanan, and Srinivasa Narasimhan. 4d visualization of dynamic events from unconstrained multi-view videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5366–5375, 2020.

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

[5] Michael Broxton, John Flynn, Ryan Overbeck, Daniel Erickson, Peter Hedman, Matthew Duvall, Jason Dourgarian, Jay Busch, Matt Whalen, and Paul Debevec. Immersive light field video with a layered mesh representation. *ACM Transactions on Graphics (TOG)*, 39(4):86–1, 2020.

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

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

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

[9] Jaesung Choe, Sunghoon Im, Francois Rameau, Minjun Kang, and In So Kweon. Volumefusion: Deep depth fusion for 3d scene reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16086–16095, 2021.

[10] Jaesung Choe, Kyungdon Joo, Tooba Imtiaz, and In So Kweon. Volumetric propagation network: Stereo-lidar fusion for long-range depth estimation. *IEEE Robotics and Automation Letters*, 6(3):4672–4679, 2021.

[11] Jaesung Choe, Byeongin Joung, Francois Rameau, Jaesik Park, and In So Kweon. Deep point cloud reconstruction. *arXiv preprint arXiv:2111.11704*, 2021.

[12] Gene Chou, Ilya Chugunov, and Felix Heide. Gensdf: Two-stage learning of generalizable signed distance functions. *Advances in Neural Information Processing Systems*, 35:24905–24919, 2022.

[13] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (ToG)*, 34(4):1–13, 2015.

[14] Paul Debevec, Yizhou Yu, and George Borshukov. Efficient view-dependent image-based rendering with projective texture-mapping. In *Eurographics Workshop on Rendering Techniques*, pages 105–116. Springer, 1998.

[15] Manfredo P Do Carmo. *Differential geometry of curves and surfaces: revised and updated second edition*. Courier Dover Publications, 2016.

[16] Yilun Du, Yinan Zhang, Hong-Xing Yu, Joshua B Tenenbaum, and Jiajun Wu. Neural radiance flow for 4d view synthesis and video processing. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14304–14314. IEEE Computer Society, 2021.

[17] Arda Duzceker, Silvano Galliani, Christoph Vogel, Pablo Speciale, Mihai Dusmanu, and Marc Pollefeys. Deepvideomvs: Multi-view stereo on video with recurrent spatio-temporal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15324–15333, 2021.

[18] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017.

[19] Jiemin Fang, Taoran Yi, Xinggang Wang, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Matthias Nießner, and Qi Tian. Fast dynamic radiance fields with time-aware neural voxels. *arXiv preprint arXiv:2205.15285*, 2022.

[20] Qiancheng Fu, Qingshan Xu, Yew-Soon Ong, and Wenbing Tao. Geo-neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction. *arXiv preprint arXiv:2205.15848*, 2022.

[21] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5712–5721, 2021.

[22] Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Monocular dynamic view synthesis: A reality check. *arXiv preprint arXiv:2210.13445*, 2022.

[23] Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and Thomas Funkhouser. Local deep implicit functions for 3d shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4857–4866, 2020.

[24] Steven J Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F Cohen. The lumigraph. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 43–54, 1996.

[25] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 2020.

[26] Haoyu Guo, Sida Peng, Haotong Lin, Qianqian Wang, Guofeng Zhang, Hujun Bao, and Xiaowei Zhou. Neural 3d scene reconstruction with the manhattan-world assumption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5511–5520, 2022.

[27] Sunghoon Im, Hae-Gon Jeon, Stephen Lin, and In So Kweon. Dpsnet: End-to-end deep plane sweep stereo. *arXiv preprint arXiv:1905.00538*, 2019.

[28] Olga Karpenko, John F Hughes, and Ramesh Raskar. Free-form sketching with variational implicit surfaces. In *Computer Graphics Forum*, volume 21, pages 585–594. Wiley Online Library, 2002.

[29] Mijeong Kim, Seonguk Seo, and Bohyung Han. Infonerf: Ray entropy minimization for few-shot neural volume rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12912–12921, 2022.

[30] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017.

[31] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1611–1621, 2021.

[32] Marc Levoy and Pat Hanrahan. Light field rendering. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 31–42, 1996.

[33] Tzu-Mao Li, Miika Aittala, Frédo Durand, and Jaakko Lehtinen. Differentiable monte carlo ray tracing through edge sampling. *ACM Transactions on Graphics (TOG)*, 37(6):1–11, 2018.

[34] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6498–6508, 2021.

[35] Jia-Wei Liu, Yan-Pei Cao, Weijia Mao, Wenqiao Zhang, David Junhao Zhang, Jussi Keppo, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Devrf: Fast deformable voxel radiance fields for dynamic scenes. *arXiv preprint arXiv:2205.15723*, 2022.

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

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

[38] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987.

[39] Baorui Ma, Zhizhong Han, Yu-Shen Liu, and Matthias Zwicker. Neural-pull: Learning signed distance functions from point clouds by learning to pull space onto surfaces. *arXiv preprint arXiv:2011.13495*, 2020.

[40] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019.

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

[42] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *arXiv preprint arXiv:2201.05989*, 2022.

[43] Zak Murez, Tarrence van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. Atlas: End-to-end 3d scene reconstruction from posed images. In *European conference on computer vision*, pages 414–431. Springer, 2020.

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

[45] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Occupancy flow: 4d reconstruction by learning particle dynamics. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5379–5389, 2019.

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

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

[48] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L Davidson, Sameh Khamis, Mingsong Dou, et al. Holoportation: Virtual 3d teleportation in real-time. In *Proceedings of the 29th annual symposium on user interface software and technology*, pages 741–754, 2016.

[49] Stanley Osher, Ronald Fedkiw, and K Piechor. Level set methods and dynamic implicit surfaces. *Appl. Mech. Rev.*, 57(3):B15–B15, 2004.

[50] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019.

[51] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021.

[52] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph.*, 40(6), dec 2021.

[53] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *European Conference on Computer Vision*, pages 523–540. Springer, 2020.

[54] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021.

[55] Yi-Ling Qiao, Alexander Gao, and Ming Lin. Neuphysics: Editable neural geometry and physics from monocular videos. In *Advances in Neural Information Processing Systems*, 2022.

[56] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[57] Yufan Ren, Tong Zhang, Marc Pollefeys, Sabine Süsstrunk, and Fangjinhua Wang. Volrecon: Volume rendering of signed ray distance functions for generalizable multi-view reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16685–16695, 2023.

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

[59] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2437–2446, 2019.

[60] Miroslava Slavcheva, Maximilian Baust, Daniel Cremers, and Slobodan Ilic. Killingfusion: Non-rigid 3d reconstruction without correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1386–1395, 2017.

[61] Liangchen Song, Anpei Chen, Zhong Li, Zhang Chen, Lele Chen, Junsong Yuan, Yi Xu, and Andreas Geiger. Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields. *arXiv preprint arXiv:2210.15947*, 2022.

[62] Olga Sorkine and Marc Alexa. As-rigid-as-possible surface modeling. In *Symposium on Geometry processing*, volume 4, pages 109–116, 2007.

[63] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. *arXiv preprint arXiv:2111.11215*, 2021.

[64] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15598–15607, 2021.

[65] Jiaxiang Tang. Torch-ngp: a pytorch implementation of instant-ngp, 2022. https://github.com/ashawkey/torch-ngp.

[66] Demetri Terzopoulos, John Platt, Alan Barr, and Kurt Fleischer. Elastically deformable models. In *Proceedings of the 14th annual conference on Computer graphics and interactive techniques*, pages 205–214, 1987.

[67] Itsuki Ueda, Yoshihiro Fukuhara, Hirokatsu Kataoka, Hiroaki Aizawa, Hidehiko Shishido, and Itaru Kitahara. Neural density-distance fields. *arXiv preprint arXiv:2207.14455*, 2022.

[68] Michael Waechter, Nils Moehrle, and Michael Goesele. Let there be color! large-scale texturing of 3d reconstructions. In *European conference on computer vision*, pages 836–850. Springer, 2014.

[69] Chaoyang Wang, Ben Eckart, Simon Lucey, and Orazio Gallo. Neural trajectory fields for dynamic novel view synthesis. *arXiv preprint arXiv:2105.05994*, 2021.

[70] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021.

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

[72] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9421–9431, 2021.

[73] Guandao Yang, Serge Belongie, Bharath Hariharan, and Vladlen Koltun. Geometry processing with neural fields. *Advances in Neural Information Processing Systems*, 34:22483–22497, 2021.

[74] Gengshan Yang, Minh Vo, Neverova Natalia, Deva Ramanan, Vedaldi Andrea, and Joo Hanbyul. Banmo: Building animatable 3d neural models from many casual videos. In *CVPR*, 2022.

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

[76] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *NeurIPS*, 34:4805–4815, 2021.

[77] Lior Yariv, Peter Hedman, Christian Reiser, Dor Verbin, Pratul P Srinivasan, Richard Szeliski, Jonathan T Barron, and Ben Mildenhall. Bakedsdf: Meshing neural sdfs for real-time view synthesis. *arXiv preprint arXiv:2302.14859*, 2023.

[78] Jae Shin Yoon, Kihwan Kim, Orazio Gallo, Hyun Soo Park, and Jan Kautz. Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5336–5345, 2020.

[79] Alex Yu, Sara Fridovich-Keil, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. *arXiv preprint arXiv:2112.05131*, 2021.

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

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

[82] C Lawrence Zitnick, Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High-quality video view interpolation using a layered representation. *ACM transactions on graphics (TOG)*, 23(3):600–608, 2004.