HelixSurf: A Robust and Efficient Neural Implicit Surface Learning of Indoor Scenes with Iterative Intertwined Regularization

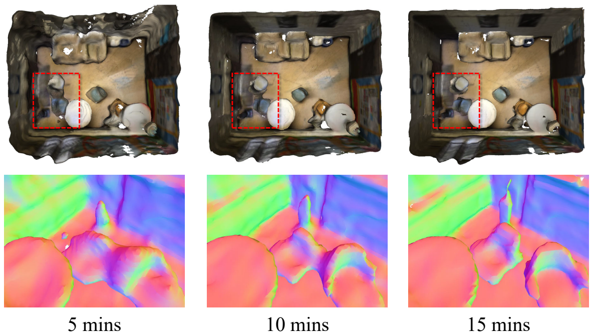
Zhihao Liang[1,*], Zhangjin Huang[1,*], Changxing Ding[1] and Kui Jia[1,2,†]
[1]South China University of Technology, [2]Peng Cheng Laboratory
{eezhihaoliang, eehuangzhangjin}@mail.scut.edu.cn, {chxding, kuijia}@scut.edu.cn,

Abstract

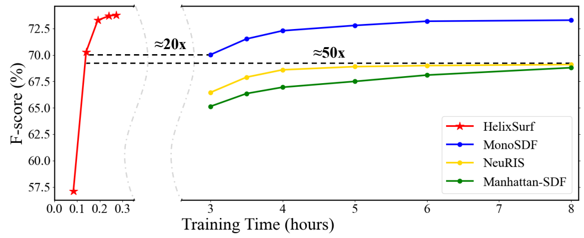
Recovery of an underlying scene geometry from multi-view images stands as a long-time challenge in computer vision research. The recent promise leverages neural implicit surface learning and differentiable volume rendering, and achieves both the recovery of scene geometry and synthesis of novel views, where deep priors of neural models are used as an inductive smoothness bias. While promising for object-level surfaces, these methods suffer when coping with complex scene surfaces. In the meanwhile, traditional multi-view stereo can recover the geometry of scenes with rich textures, by globally optimizing the local, pixel-wise correspondences across multiple views. We are thus motivated to make use of the complementary benefits from the two strategies, and propose a method termed Helix-shaped neural implicit Surface learning or HelixSurf; HelixSurf uses the intermediate prediction from one strategy as the guidance to regularize the learning of the other one, and conducts such intertwined regularization iteratively during the learning process. We also propose an efficient scheme for differentiable volume rendering in HelixSurf. Experiments on surface reconstruction of indoor scenes show that our method compares favorably with existing methods and is orders of magnitude faster, even when some of existing methods are assisted with auxiliary training data. The source code is available at https://github.com/Gorilla-Lab-SCUT/HelixSurf.

1. Introduction

Surface reconstruction of a scene from a set of observed multi-view images stands as a long-term challenge in computer vision research. A rich literature [4, 11, 15] exists to address the challenge, including different paradigms of methods from stereo matching to volumetric fusion.



(a) Results of HelixSurf on an example scene from ScanNet at three training checkpoints, where we use color codes to visualize surface normals.



(b) Training curves of different methods on ScanNet. The empirical training time of each method is measured on a machine with a single NVIDIA RTX 3090 GPU.

Figure 1. Efficacy and efficiency of our proposed HelixSurf.

Among them, the representative methods of multi-view stereo (MVS) [12,37,44,52] first recover the properties (e.g, depth and/or normal) of discrete surface points, by globally optimizing the local, pixel-wise correspondences across the multi-view images, where photometric and geometric consistencies across views are used as the optimization cues, and a continuous fitting method (e.g., Poisson reconstruction [17, 18]) is then applied to recover a complete surface. MVS methods usually make a reliable recovery only on surface areas with rich textures.

More recently, differentiable volume rendering is proposed that connects the observed multi-view images with neural modeling of the implicit surface and radiance field [28, 41, 48]. They show a surprisingly good promise for recovery of object-level surfaces, especially when the object

* indicates equal contribution.
† correspondence to Kui Jia <kuijia@scut.edu.cn>.

masks are available in the observed images [24, 49]; indeed, these methods favor a continuous, closed surface given that a single deep network is used to model the scene space, whose deep prior induces a smoothness bias for surface recovery [41, 48]. For complex scene surfaces, however, the induced smoothness bias is less capable to regularize the learning and recover the scene surface with fine geometry details [34, 51].

To overcome the limitation, we observe that the strategies from the two paradigms of MVS and neural implicit learning are different but potentially complementary to the task. We are thus motivated to make use of the complementary benefits with an integrated solution. In this work, we achieve the goal technically by using the intermediate prediction from one strategy as the guidance to regularize the learning/optimization of the other one, and conducting such *intertwined regularization iteratively* during the process. Considering that the iterative intertwined regularization makes the optimization curve as a shape of double helix, we term our method as *Helix-shaped neural implicit Surface learning or HelixSurf*. Given that MVS predictions are less reliable for textureless surface areas, we regularize the learning on such areas in HelixSurf by leveraging the homogeneity inside individual superpixels of observed images. We also improve the efficiency of differentiable volume rendering in HelixSurf, by maintaining dynamic occupancy grids that can adaptively guide the point sampling along rays; our scheme improves the learning efficiency with orders of magnitude when compared with existing neural implicit surface learning methods, even with the inclusion of MVS inference time. An illustration of the proposed HelixSurf is given in Fig. 2. Experiments on the benchmark datasets of ScanNet [6] and Tanks and Temples [20] show that our method compares favorably with existing methods, and is orders of magnitude faster. We note that a few recent methods [40, 50] use geometric cues provided by models pre-trained on auxiliary data to regularize the neural implicit surface learning; compared with them, our method achieves better results as well. Our technical contributions are summarized as follows.

- We present a novel method of *HelixSurf* for reconstruction of indoor scene surface from multi-view images. HelixSurf enjoys the complementary benefits of the traditional MVS and the recent neural implicit surface learning, by regularizing the learning/optimization of one strategy iteratively using the intermediate prediction from the other;
- MVS methods make less reliable predictions on textureless surface areas. We further devise a scheme that regularizes the learning on such areas by leveraging the region-wise homogeneity organized by superpixels in each observed image.

2. Related Works

2.1. PatchMatch based Multi-view Stereo

3D reconstruction from posed multi-view images is a fundamental but challenging task in computer vision. Among all the techniques in the literature, PatchMatch based Multi-view Stereo (PM-MVS) is traditionally the most explored one [11, 15]. PM-MVS methods [12, 35–38, 44, 52] represent the geometric with depth and/or normal maps. They estimate depth and/or normal of each pixel by exploiting inter-image photometric and geometric consistency and then fuse all the depth maps into a global point cloud with filtering operations, which can be subsequently processed using meshing algorithms [17, 21], *e.g.* Screened Poisson surface reconstruction [18], to recover complete surface. These traditional methods have achieved great success on various occasions and can produce plausible geometry of textured surfaces, but there exist artifacts and missing parts in the areas without rich textures. Indeed, their optimization highly relies on the photometric measure to discriminate which random estimate is the best guess. In the case of indoor scenes with textureless areas [35, 44], the inherent homogeneity inactivates the photometric measure and consequently poses difficulties to the accurate depth estimation. With the development of deep learning, learning-based MVS methods [16, 39, 45–47] demonstrate promising performance in recent years. However, they crucially rely on ground-truth 3D data for supervision, which hinders their practical application.

2.2. Neural Implicit Surface

In contrast to classic explicit representation, recent works [5, 27, 32] implicitly represent surfaces via learning neural networks, which models continuous surface with Multi-Layer Perceptron (MLP) and makes it more feasible and efficient to represent complex geometries with arbitrary typologies. For the task of multi-view reconstruction, the 3D geometry is represented by a neural network that outputs either a signed/unsigned distance field or an occupancy field. Some works [24, 30, 49] utilize surface rendering to enable the reconstruction of 3D shapes from 2D images, but they always rely on extra object masks. Inspired by the success of NeRF [28], recent works [31, 41, 48] attach differentiable volume rendering techniques to reconstruction, which eliminates the need of mask and achieves impressive reconstruction. And follow-up works [8, 10, 42] further improve the geometry quality with fine-grained surface details. Although these methods show better accuracy and completeness compared with the traditional MVS methods, they still suffer from the induced smoothness bias of deep network [34, 51], which discourages them to regularize the learning and recover fine details in scene reconstruction. Most recent works [14, 40, 50] try to get rid of this dilemma by incorporating geometric cues provided by models pre-trained on auxiliary data. Our HelixSurf integrates traditional PM-MVS and neural implicit learning surface in complementary mechanisms and achieves better results than these methods.

3. Preliminary

In this section, we give technical backgrounds and math notations that are necessary for presentation of our proposed method in subsequent sections.

Neural Implicit Surface Representation Among the choices of neural implicit surface representation [5, 27, 32], we adopt DeepSDF [32] that learns to encode a continuous surface as the zero-level set of a signed distance field (SDF) $f : \mathbb{R}^3 \to \mathbb{R}$, which is typically parameterized as an MLP; for any point $\boldsymbol{x} \in \mathbb{R}^3$ in the 3D space, $|f(\boldsymbol{x})|$ assigns its distance to the surface $\mathcal{S} = \left\{\boldsymbol{x} \in \mathbb{R}^3 | f(\boldsymbol{x}) = 0\right\}$; by convention, we have $f(\boldsymbol{x}) < 0$ for points inside the surface and $f(\boldsymbol{x}) > 0$ for those outside.

SDF-induced Volume Rendering Differentiable volume rendering is used in NeRF [28] for synthesis of novel views. Denote a ray emanating from a viewing camera as $\boldsymbol{r}(t) = \boldsymbol{o} + t\boldsymbol{v}$, $t \geq 0$, where $\boldsymbol{o} \in \mathbb{R}^3$ is the camera center and $\boldsymbol{v} \in \mathbb{R}^3, \|\boldsymbol{v}\| = 1$ denotes the unit vector of viewing direction. NeRF models a continuous scene space as a neural radiance field $\boldsymbol{F} : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}_+ \times \mathbb{R}^3$, which for any space point $\boldsymbol{x}$ and direction $\boldsymbol{v}$, assigns $\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{v}) = (\sigma, \boldsymbol{c})$, where $\sigma \in \mathbb{R}_+$ represents the volume density at the location $\boldsymbol{x}$, and $\boldsymbol{c} \in \mathbb{R}^3$ is the view-dependent color from $\boldsymbol{x}$ along the ray $-\boldsymbol{r}$ towards $\boldsymbol{o}$. Assume N points are sampled along $\boldsymbol{r}$; the color accumulated along the ray $\boldsymbol{r}$ can be approximated, using the quadrature rule [26], as

$$\boldsymbol{C}(\boldsymbol{r}) = \sum_{i=1}^{N} T_i \alpha_i \boldsymbol{c}(\boldsymbol{r}(t_i), \boldsymbol{v}), \quad T_i = \prod_{j=1}^{i-1} (1 - \alpha_j), \quad (1)$$

where $\alpha_i = 1 - \exp(-\int_{t_i}^{t_{i+1}} \sigma(\boldsymbol{r}(t)) dt)$ denotes the opacity of a segment. While the volume density $\sigma : \mathbb{R}^3 \to \mathbb{R}_+$ is learned as a direct output of the MLP based radiance field function $\boldsymbol{F}$ in [28], it is shown in VolSDF [48] and NeuS [41] that σ can be modeled as a transformed function of the implicit SDF function f, enabling better recovery of the underlying geometry. In this work, we follow [41] to model σ as an SDF-induced volume density. With such an SDF-induced, differentiable volume rendering, the geometry f and color $\boldsymbol{c}$ can be learned by minimizing the difference between rendering results and multiple views of input images. Note that analogous to (1), the depth d of the surface from the camera center $\boldsymbol{o}$ can be approximated along the ray $\boldsymbol{r}$ as well, giving rise to

$$d(\boldsymbol{r}) = \sum_{i=1}^{N} T_i \alpha_i t_i, \quad \boldsymbol{n}(\boldsymbol{r}) = \nabla f(\boldsymbol{o} + d(\boldsymbol{r})\boldsymbol{v}), \quad (2)$$

where $\boldsymbol{n}(\boldsymbol{r}) \in \mathbb{R}^3$ denotes the surface normal at the intersection point and $\nabla f(\boldsymbol{x})$ is the gradient of SDF at $\boldsymbol{x}$.

Multi-View Stereo with PatchMatch Assume that a reference image $\boldsymbol{I}^{\text{ref}}$ and a set of source images $\mathcal{I}^{\text{src}} = \{\boldsymbol{I}^m | m = 1 \dots M\}$ capture a common scene; we write collectively as $\mathcal{I} = \{\boldsymbol{I}^{\text{ref}}, \mathcal{I}^{\text{src}}\}$. PatchMatch based multi-view stereo (PM-MVS) methods [12, 37, 44, 52] aim to recover the scene geometry by predicting the depth $d_l \in \mathbb{R}^+$ and normal $\boldsymbol{n}_l \in \mathbb{R}^3, \|\boldsymbol{n}_l\| = 1$ for each pixel in $\boldsymbol{I}^{\text{ref}}$, which is indexed by l with $l \in \{1, \dots, L\}$. Considering that any l^{th} pixel in $\boldsymbol{I}^{\text{ref}}$ may not be visible in all images in $\mathcal{I}^{\text{src}}$, the methods then predict an occlusion indicator $\mathcal{Z}^{src} = \{Z_l^m | l = 1, \dots, L, m = 1, \dots, M\}$ for $\boldsymbol{I}^{\text{ref}}$. Optimization of $\{d_l\}_{l=1}^L$, $\{\boldsymbol{n}_l\}_{l=1}^L$, and $\mathcal{Z}^{src}$ is based on enforcing photometric and geometric consistencies between corresponding patches in $\boldsymbol{I}^{\text{ref}}$ and $\mathcal{I}^{\text{src}}$; this is mathematically formulated as a probabilistic graphical model and is solved via generalized expectation-maximization (GEM) algorithm [12, 37], where PatchMatch [2, 3] is used to efficiently establish pixel-wise correspondences across multi-view images. More specifically, let $\mathcal{A}^{src} = \{\boldsymbol{A}_l^m | l = 1, \dots, L, m = 1, \dots, M\}$ denote the set of homography-warped patches from source images [38], the PatchMatch based methods optimize d_l and $\boldsymbol{n}_l$ for a pixel in the reference image as

$$\begin{aligned} \{d_l^*, \boldsymbol{n}_l^*\} &= \arg\max P(d_l, \boldsymbol{n}_l | \mathcal{A}^{src}, \mathcal{Z}^{src}) \\ &\propto \arg\max P(\mathcal{A}^{src} | d_l, \boldsymbol{n}_l, \mathcal{Z}^{src}) P(d_l, \boldsymbol{n}_l) \\ &= \arg\min \sum_{m=1}^{M} P_l(m) \xi_l^m(d_l, \boldsymbol{n}_l) \\ \text{with } \xi_l^m &= 1 - \rho_l^m(d_l, \boldsymbol{n}_l) + \eta \min(\varphi_l^m(d_l, \boldsymbol{n}_l), \varphi_{\max}), \end{aligned} \quad (3)$$

where $\rho_l^m(d_l, \boldsymbol{n}_l)$ denotes the color similarity between the reference patch $\boldsymbol{A}_l^{\text{ref}}$ and source patch $\boldsymbol{A}_l^m$ based on normalized cross-correlation, which is a function of d_l and $\boldsymbol{n}_l$, and $\varphi_l^m(d_l, \boldsymbol{n}_l)$ is the forward-backward reprojection error to evaluate the geometric consistency incurred by the predicted d_l and $\boldsymbol{n}_l$, which is capped by a pre-defined $\varphi_{\max}$; the probability $P_l(m)$ serves for view selection that assigns different weights to the M source images. Indeed, source images with small values of $P_l(m)$ are less informative; hence Monte-Carlo view sampling is used in [52] to draw samples according to $P_l(m)$. Assume that the selected views form a subset $S \subset \{1 \dots M\}$, the problem (3) can be simplified as

$$\{d_l^*, \boldsymbol{n}_l^*\} = \arg\min \frac{1}{|S|} \sum_{m \in S} \xi_l^m(d_l, \boldsymbol{n}_l). \quad (4)$$

4. HelixSurf for Intertwined Regularization of Neural Implicit Surface Learning

Given a set of calibrated RGB images $\{\boldsymbol{I}_m\}_{m=1}^M$ of an indoor scene captured from multiple views, the task is to

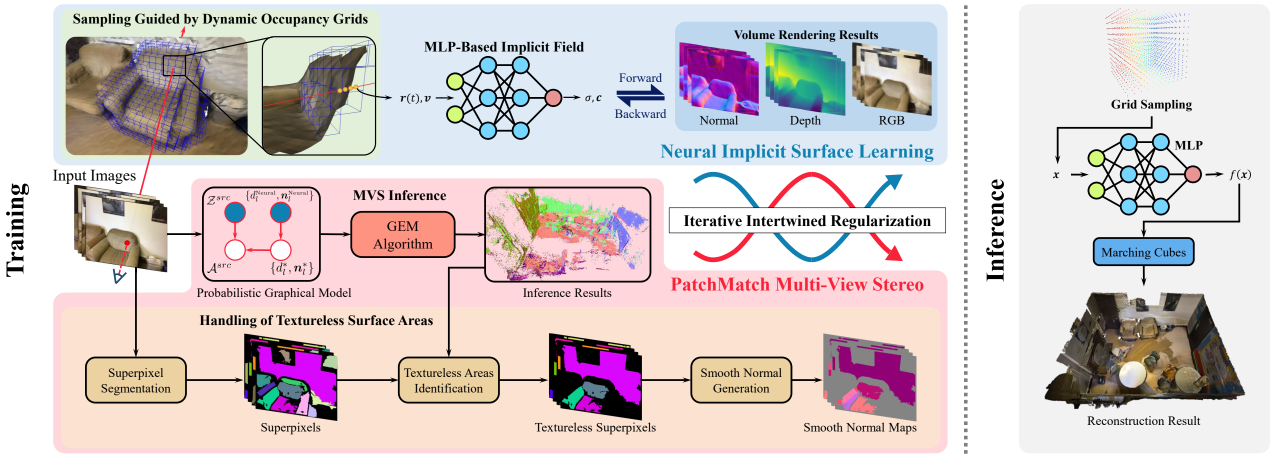


Figure 2. **Overview of HelixSurf: Helix-shaped neural implicit Surface learning**. HelixSurf integrates the neural implicit surface learning (*cf*. Section 4.1) and PatchMatch based MVS (*cf*. Section 4.2) in a robust and efficient manner. We optimize HelixSurf with an iterative intertwined regularization, which uses the intermediate prediction from one strategy as guidance to regularize the learning/optimization of the other one; given that MVS predictions are less reliable for textureless surface areas, we additionally devise a scheme that regularizes the learning on such areas by leveraging the homogeneity per superpixel in observed multi-view images (*cf*. Section 4.1.1). We also propose a scheme for point sampling along rays (*cf*. Section 4.3), which significantly improves the efficiency. At the inference stage of HelixSurf, we conduct grid sampling to query the learned SDF values at sampled points and run Marching Cubes to get the reconstruction results.

reconstruct the scene geometry with fine details. Under the framework of neural differentiable volume rendering, the task translates as learning an MLP based radiance field function $\boldsymbol{F}$ that connects the underlying scene geometry with the image observations $\{\boldsymbol{I}_m\}_{m=1}^M$; with the use of an SDF-induced volume density $\sigma(f)$, the scene surface can be reconstructed by extracting the zero-level set of the learned SDF f. As stated in Section 1, although the supervision from $\{\boldsymbol{I}_m\}_{m=1}^M$ is conducted in a pixel-wise, independent manner, the MLP based function f has *deep priors* that induce the function learning biased towards encoding continuous and piece-wise, smooth surface [41, 48]; indeed, assuming a successful learning of a ReLU-based MLP f, its zero-level set can be *exactly* recovered as a continuous polygon mesh [22]. In the meanwhile, PatchMatch based MVS methods couple the predictions of $\{d_l, \boldsymbol{n}_l\}$ for individual pixels in a probabilistic framework, and conduct the optimization *globally* such that the predicted $\{d_l, \boldsymbol{n}_l\}$ achieves an overall best consistencies of photometry and geometry across $\{\boldsymbol{I}_m\}_{m=1}^M$; after obtaining $\{d_l, \boldsymbol{n}_l\}$, a continuous, watertight surface can be fitted using Poisson reconstruction [17, 18]. The above two strategies reconstruct the surface using different but potentially complementary mechanisms. We are thus motivated to propose an integrated solution that can take both advantages of them. In this work, we achieve the goal technically by using the intermediate prediction from one strategy as the guidance to regularize the learning of the other one, and conducting such *intertwined regularization iteratively* during the learning process. Considering that the iterative intertwined regularization makes the optimization curve as a shape of double helix, we term our method as *Helix-shaped neural implicit Surface learning or HelixSurf.* Details of HelixSurf are presented as follows. An illustration is given in Fig. 2.

4.1. Regularization of Neural Implicit Surface Learning from MVS predictions

Given the set of multi-view images $\mathcal{I} = \{\boldsymbol{I}^{\text{ref}}, \mathcal{I}^{\text{src}}\}$, neural implicit surface learning via differentiable volume rendering samples rays in the 3D space; for any sampled ray $\boldsymbol{r}(t) = \boldsymbol{o} + t\boldsymbol{v}$, $t \geq 0$, in a viewing direction $\boldsymbol{v}$, assume that it emanates from the camera center $\boldsymbol{o}$ and passes through a pixel $\boldsymbol{a} \in \mathbb{R}^3$ in an image $\boldsymbol{I}$ in $\mathcal{I}$. Let $\boldsymbol{F}$ be the SDF-induced neural radiance field that models the scene geometry via the SDF function f; we can then write as $\boldsymbol{F}(\boldsymbol{r}(t), \boldsymbol{v}; f) = (\sigma(f(\boldsymbol{r}(t))), \boldsymbol{c}(\boldsymbol{r}(t), \boldsymbol{v}))$ for any point t along $\boldsymbol{r}(t)$. According to (1) of approximated volume rendering, the color $\boldsymbol{C}(\boldsymbol{r})$ accumulated along the ray $\boldsymbol{r}$ can be computed, given $\{\sigma(f(\boldsymbol{r}(t_i))), \boldsymbol{c}(\boldsymbol{r}(t_i), \boldsymbol{v})\}_{i=1}^N$ at N sampled points; the following loss defines the color based image supervision from ray $\boldsymbol{r}$ for learning $\boldsymbol{F}$ (i.e., learning the MLPs f and $\boldsymbol{c}$, see Section 3 for the details):

$$\mathcal{L}_{\text{Neural}}(\boldsymbol{r}; f, \boldsymbol{c}) = \texttt{SmoothL1}(\boldsymbol{C}(\boldsymbol{r}; f, \boldsymbol{c}), \boldsymbol{a}(\boldsymbol{r})). \tag{5}$$

We can also compute the depth $d(\boldsymbol{r}; f)$ and surface normal $\boldsymbol{n}(\boldsymbol{r}; f)$ according to (2).

Section 3 suggests that given $\mathcal{I}$, PatchMatch based MVS methods can predict pairs of depth and surface normal for pixels in the observed reference image. Such methods usually produce a sparse set of predictions on texture-rich surface areas [37, 43]. Without loss of generality, assume that $\{d_{\boldsymbol{a}}^{\text{MVS}}, \boldsymbol{n}_{\boldsymbol{a}}^{\text{MVS}}\}$ are the MVS prediction for the pixel $\boldsymbol{a}$ in the image $\boldsymbol{I}$. We use $\{d_{\boldsymbol{a}}^{\text{MVS}}, \boldsymbol{n}_{\boldsymbol{a}}^{\text{MVS}}\}$ to regularize the learning of f in the current iteration, based on the following loss

$$\mathcal{L}_{\text{MVSRegu}}(\boldsymbol{r}; f) = w(\boldsymbol{r}) \left(\left| d(\boldsymbol{r}; f) - d_{\boldsymbol{a}}^{\text{MVS}} \right| + \left| \boldsymbol{n}(\boldsymbol{r}; f) - \boldsymbol{n}_{\boldsymbol{a}}^{\text{MVS}} \right| \right), \\ \text{with } \ w(\boldsymbol{r}) = \mathbb{1}_{\text{MVSRegu}}(\boldsymbol{r}) \cdot (1 - \left| \boldsymbol{C}(\boldsymbol{r}) - \boldsymbol{a}(\boldsymbol{r}) \right|) \tag{6}$$

where $\mathbb{1}_{\text{MVSRegu}}(\boldsymbol{r})$ is an indicator to cope with the case when $\{d_{\boldsymbol{a}}^{\text{MVS}}, \boldsymbol{n}_{\boldsymbol{a}}^{\text{MVS}}\}$ are not predicted by MVS for the pixel $\boldsymbol{a}$.

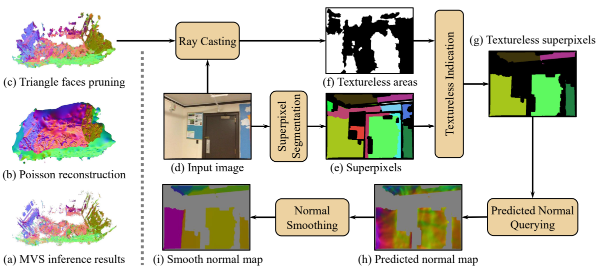


Figure 3. Illustration of handling textureless surface areas. (a): the inference results of PM-MVS, (b): watertight surface mesh $\mathcal{M}^{\text{MVS}}$ reconstructed by Poisson reconstruction from (a), (c): surface $\mathcal{M}_{-}^{\text{MVS}}$ obtained by pruning textureless triangle faces from (b), (d): an input image, (e): superpixels extracted by the graph-based segmentation algorithm [9], (f): textureless areas obtained by ray casting, (g): superpixels covered by textureless areas, (h): surface normal map predicted by HelixSurf, (i): smooth normal map obtained by aggregating the normals in textureless superpixels.

4.1.1 Handling of Textureless Surface Areas

PatchMatch based MVS methods make reliable predictions only on texture-rich surface areas. We resort to other sources to regularize the neural implicit learning for textureless surface areas. Our motivation is based on the observation that textureless surface areas tend to be both homogeneous in color and geometrically smooth; indeed, when the surface areas are of high curvature or when they have different colors, 2D image projections of such areas would have richer textures. The projected 2D image counterparts of textureless surface areas in fact correspond to those in images that can be organized as superpixels. We thus propose to further regularize the neural implicit surface learning by leveraging the homogeneity of image superpixels.

We technically encourage the predicted normals of surface points, whose 2D projections fall in a same superpixel, to be close. For any image $\boldsymbol{I}$ in $\mathcal{I}$, we pre-compute its region partitions of superpixels using methods such as [1, 9]. Let $\tilde{\boldsymbol{r}}$ be a ray passing through a pixel $\tilde{\boldsymbol{a}}$ that falls in a superpixel of $\boldsymbol{I}$; denote the superpixel as $\tilde{\boldsymbol{A}}_{\tilde{\boldsymbol{a}}}$. We know that the volume rendering in HelixSurf predicts surface normal $\boldsymbol{n}(\tilde{\boldsymbol{r}}; f)$ for the ray $\tilde{\boldsymbol{r}}$, and denote as $\{\boldsymbol{n}'(\tilde{\boldsymbol{r}}'; f) | \tilde{\boldsymbol{a}}' \in \tilde{\boldsymbol{A}}_{\tilde{\boldsymbol{a}}}\}$ the predicted surface normals for all pixels in $\tilde{\boldsymbol{A}}_{\tilde{\boldsymbol{a}}}$. We first compute $\boldsymbol{n}_{\tilde{\boldsymbol{a}}, \boldsymbol{I}}^{\text{Smooth}} = \sum_{i=1}^{|\tilde{\boldsymbol{A}}_{\tilde{\boldsymbol{a}}}|} \boldsymbol{n}'_i / |\tilde{\boldsymbol{A}}_{\tilde{\boldsymbol{a}}}|$, and then apply the above computation to all those images in $\mathcal{I}$ that capture the same surface point and have the corresponding pixels of $\tilde{\boldsymbol{a}}$ in $\boldsymbol{I}$. Assume we have a total of M' such image, we compute $\boldsymbol{n}_{\tilde{\boldsymbol{a}}}^{\text{Smooth}} = \sum_{m=1}^{M'+1} \boldsymbol{n}_{\tilde{\boldsymbol{a}}_m, \boldsymbol{I}_m}^{\text{Smooth}} / (M'+1)$ and enforce closeness of surface normal predictions for pixels both inside a superpixel and across multi-view images with the following loss

$$\mathcal{L}_{\text{Smooth}}(\tilde{\boldsymbol{r}}; f) = \mathbb{1}_{\text{Smooth}}(\tilde{\boldsymbol{r}}) \cdot \left| \boldsymbol{n}(\tilde{\boldsymbol{r}}; f) - \boldsymbol{n}_{\tilde{\boldsymbol{a}}}^{\text{Smooth}} \right|, \tag{7}$$

where $\mathbb{1}_{\text{Smooth}}(\tilde{\boldsymbol{r}})$ indicates whether the pixel $\tilde{\boldsymbol{a}}$ cast by the ray $\tilde{\boldsymbol{r}}$ belongs to a textureless area. In practice, we identify the textureless areas for a surface $\mathcal{S} = \left\{ \boldsymbol{x} \in \mathbb{R}^3 | f(\boldsymbol{x}) = 0 \right\}$ as follows. We first use MVS methods to produce a sparse set of depth and normal predictions, to which we apply Poisson reconstruction [17, 18] and obtain a watertight surface mesh $\mathcal{M}^{\text{MVS}}$ (Fig. 3(b)). We prune those triangle faces in $\mathcal{M}^{\text{MVS}}$ that contain no the depths and normals predicted by the MVS methods, resulting in $\mathcal{M}_{-}^{\text{MVS}}$. For an image $\boldsymbol{I}$, we conduct ray casting and treat the pixels whose associated rays do not hit $\mathcal{M}_{-}^{\text{MVS}}$ as those belonging to textureless areas (Fig. 3(e)). The overall scheme is illustrated in Fig. 3. Please refer to the supplementary for more details.

4.2. Regularization of Multi-View Stereo from Neural Implicit Surface Learning

Eq. (3) of MVS methods optimize the depth and normal predictions by maximizing a posterior probability and a prior of $P(d, \boldsymbol{n})$ (cf. line 2 in Eq. (3)). Without other constraints, $P(d, \boldsymbol{n})$ is usually set as a uniformly random distribution. In HelixSurf, it is obviously feasible to use the depth and normal learned in the current iteration of neural implicit learning as the prior.

More specifically, given d_l, $\boldsymbol{n}_l$, $\mathcal{A}^{src}$, and $\mathcal{Z}^{src}$ denoted as in Section 3, let d_l^{Neural} and $\boldsymbol{n}_l^{\text{Neural}}$ be the depth and normal learned in the current iteration of neural implicit learning for the corresponding pixel in an observed image. We can improve MVS predictions using

$$\{d_l^*, \boldsymbol{n}_l^*\} = \arg\max P(d_l, \boldsymbol{n}_l | \mathcal{A}^{src}, \mathcal{Z}^{src}, d_l^{\text{Neural}}, \boldsymbol{n}_l^{\text{Neural}}) \\ \propto \arg\max P(\mathcal{A}^{src} | d_l, \boldsymbol{n}_l, \mathcal{Z}^{src}) P(d_l, \boldsymbol{n}_l | d_l^{\text{Neural}}, \boldsymbol{n}_l^{\text{Neural}}). \tag{8}$$

Qualitative results in Section 5.2 show that MVS methods with priors of a uniformly random distribution tend to produce noisy results with outliers, which would impair the iterative learning in HelixSurf. Instead, the proposed (8) gives better results.

4.3. Improving the Efficiency by Establishing Dynamic Space Occupancies

Differentiable volume rendering suffers from the heavy cost of point sampling along rays for accumulating pixel colors [28, 41, 48]. While a common coarse-to-fine sampling strategy is used in these methods, it still counts as the main computation. In this work, we are inspired by Instant-NGP [23, 29] and adopt a simple yet effective sampling scheme, which establishes dynamic occupancies in the 3D scene space and adaptively guides the point sampling along rays. Fig. 2 gives the illustration. More details of our scheme are given in the supplementary material. Fig. 1(b) shows that our scheme improves the training efficiency at orders of magnitude when compared with existing neural implicit surface learning methods.

4.4. Training and Inference

At each iteration of HelixSurf training, we randomly sample pixels from the images in $\mathcal{I}$ and define the set of camera rays passing through these pixels as $\mathcal{R} \cup \tilde{\mathcal{R}}$, where $\mathcal{R}$ and $\tilde{\mathcal{R}}$ contain rays passing through texture-rich and textureless areas respectively. We optimize the following problem to learn the MLP based functions f and $\boldsymbol{c}$

$$\begin{aligned}\min_{f,\boldsymbol{c}}(& \sum_{\boldsymbol{r}\in\mathcal{R}} \mathcal{L}_{\text{Neural}}(\boldsymbol{r}; f, \boldsymbol{c}) + \sum_{\tilde{\boldsymbol{r}}\in\tilde{\mathcal{R}}} \mathcal{L}_{\text{Neural}}(\tilde{\boldsymbol{r}}; f, \boldsymbol{c}) + \\ & \lambda_{\text{MVSRegu}} \sum_{\boldsymbol{r}\in\mathcal{R}} \mathcal{L}_{\text{MVSRegu}}(\boldsymbol{r}; f) + \lambda_{\text{Smooth}} \sum_{\tilde{\boldsymbol{r}}\in\tilde{\mathcal{R}}} \mathcal{L}_{\text{Smooth}}(\tilde{\boldsymbol{r}}; f) \\ & \lambda_{\text{Eik}} \sum_{\boldsymbol{x}\in\mathbb{R}^3} \mathcal{L}_{\text{Eik}}(\boldsymbol{x}; f)),\end{aligned} \tag{9}$$

where $\mathcal{L}_{\text{Eik}}(\boldsymbol{x}; f)$ is the Eikonal loss [13] that regularizes the learning of SDF f, and $\lambda_{\text{MVSRegu}}, \lambda_{\text{Smooth}}, \lambda_{\text{Eik}}$ are hyperparameters weighting different loss terms.

During inference, we apply marching cubes [25] algorithm to extract the underlying surface from the learned SDF f.

5. Experiments

Datasets We conduct experiments using the benchmark dataset of ScanNet [6] and Tanks and Temples [20]. ScanNet has 1613 indoor scenes with precise camera calibration parameters and surface reconstructions via the state-of-the-art SLAM technique [7]. Tanks and Temples has multiple large-scale indoor and outdoor scenes. For the ScanNet, we follow ManhattanSDF [14] and select 4 scenes to conduct our experiments. As for Tanks and Temples, we follow MonoSDF [50] to select four large-scale indoor scenes to further investigate the extensibility of HelixSurf.

Implementation Details We implement HelixSurf in PyTorch [33] framework with CUDA extensions, and customized a PM-MVS module for HelixSurf according to COLMAP [37] and ACMP [44]. We use the Adam optimizer [19] with a learning rate of 1e-3 for network training, and set $\lambda_{\text{MVSRegu}}, \lambda_{\text{Smooth}}, \lambda_{\text{Eik}}$ to 0.5, 0.01, 0.03, respectively. For each iteration, we sample 5000 rays to train the model and use customized CUDA kernels for calculating the α-compositing colors of the sampled points along each ray as Eq. (1). We train the model for a total of $24K$ iterations. To maintain dynamic occupancy grids, we update the grids after every 16 training iterations and cap the mean density of grids by 1e-2 as the density threshold τ_{occu}.

Method	Acc↓	Comp↓	Prec↑	Recall↑	F-score↑	Time↓
COLMAP [37]	0.047 🟤	0.235	0.711	0.441	0.537	*133*
ACMP [44]	0.118	0.081	0.531	0.581	0.555	*10*
NeRF [28]	0.735	0.177	0.131	0.290	0.176	$>$ 1k
VolSDF [48]	0.414	0.120	0.321	0.394	0.346	825
NeuS [41]	0.179	0.208	0.313	0.275	0.291	531
Manhattan-SDF† [14]	0.053	0.056	0.715 🟤	0.664	0.688	528
NeuRIS† [40]	0.050	0.049 🟤	0.714	0.670 🟤	0.691 🟤	406
MonoSDF† [50]	0.035 🟡	0.048 ⚪	0.799 🟡	0.681 ⚪	0.733 ⚪	708
HelixSurf	0.038 ⚪	0.044 🟡	0.786 ⚪	0.727 🟡	0.755 🟡	**33**

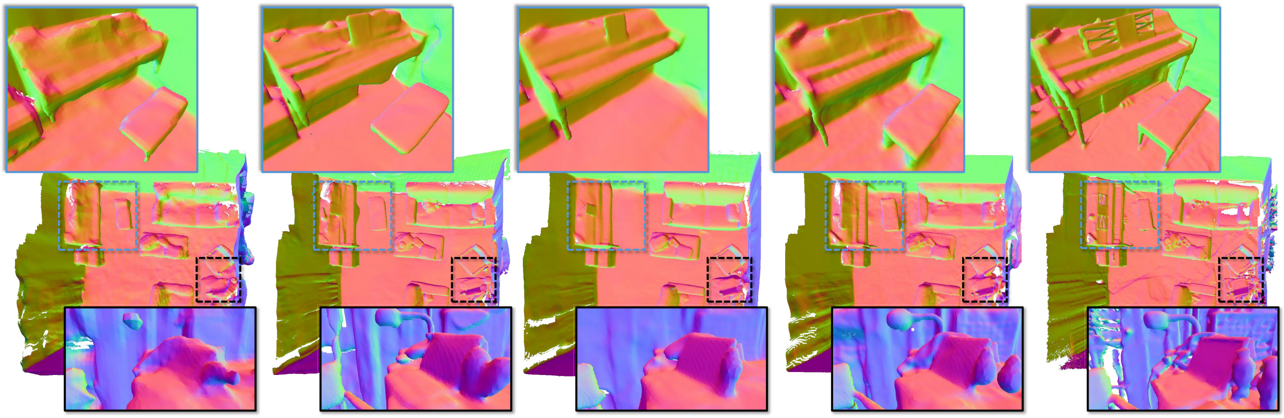
Table 1. **Reconstruction metrics comparisons on ScanNet [6].** We compare our method with the state-of-the-art neural implicit surface learning methods [14, 28, 40, 41, 48, 50] and PatchMatch based multi-view stereo methods (PM-MVS) [37, 44]. Methods marked with † are assisted with auxiliary training data, and vice versa. We mark the methods performing with least error using gold 🟡, silver ⚪, and bronze 🟤 medals. The last column shows time consumption (in minutes) for *PM-MVS* methods, Nueral implicit surface learning methods, and **HelixSurf**. Note that the time for HelixSurf includes both MVS inference and neural implicit surface learning.

Evaluation Metrics For 3D reconstruction, we assess the reconstructed surfaces in terms of Accuracy, Completeness, Precision, Recall, and F-score. To evaluate the MVS predictions, we compute the distance differences for depth maps and count the angle errors for normal maps. Please refer to the supplementary for more details about these evaluation metrics.

5.1. Comparisons

We evaluate the 3D geometry metrics and time consumption of our proposed HelixSurf against existing methods on ScanNet [6], as shown in Tab. 1. Each quantitative result is averaged over all the selected scenes. For the geometry comparison, HelixSurf manifestly surpasses existing methods in almost every metrics, even some of the comparison methods are assisted with auxiliary training data. And the qualitative results in Fig. 4 further support the quantitative analyses. Without auxiliary training data, HelixSurf is capable of handling the textureless surface areas where other methods fail to tackle, as in Fig. 4(a). Moreover, HelixSurf produces better details of objects than those methods using auxiliary training data, as in Fig. 4(b). As for learning

(a) Comparisons with existing methods when these methods do not use auxiliary training data.



(b) Comparisons with existing methods when they invoke the use of auxiliary training data. Note that our HelixSurf do not use auxiliary training data.

Figure 4. **Qualitative geometry comparisons on ScanNet.** Compared to existing methods, our method can better reconstruct the scene details (*e.g.* the lamp, the cabinet and chair handles) and the smooth regions (*e.g.* the floor and walls). Surface normals are visualized as coded colors.

time, the data in Tab. 1 indicates that HelixSurf improves the learning efficiency with orders of magnitude when compared with existing neural implicit surface learning methods, even with the inclusion of MVS inference time.

5.2. Ablation Studies

HelixSurf is optimized with interactive intertwined regularization as stated in Section 4. We design elaborate experiments to evaluate the efficacy of this regularization. Furthermore, the sampling guided by dynamic occupancy grids (*cf*. Section 4.3) is essential to realize fast training convergence. We thus compare it with the ordinary sampling alternative. These studies are conducted on the ScanNet dataset [6].

Analysis on the regularization of neural implicit surface learning from MVS predictions The MVS inference results effectively regularize the neural implicit surface learning (*cf*. Section 4.1) and facilitate the network to capture fine details. The results in Tab. 2 illustrate that the MVS predictions effectively promote the surface learning and the *regularized* MVS predictions can further improve the quality of reconstruction. Nonetheless, the MVS predictions are less reliable on the textureless surface areas, we thus leverage the homogeneity inside individual superpixels and devise a scheme (*cf*. Section 4.1.1) to regularize the learning on such areas. As shown in Fig. 5, our proposed scheme handles the textureless surface areas and reconstructs smoother surface on the basis of maintaining the details of non-planar regions.

Analysis on the regularization of MVS from neural implicit surface learning During the training process of the neural surface, the underlying geometries are progressively recovered. We use the learned depths and normals as priors to regularize MVS, which clears up the artifacts produced by the ordinary MVS and enables the double helix to forward and rise. Fig. 6 qualitatively shows the inference results of the ordinary MVS method (Fig. 6(a)) and the regularized MVS (Fig. 6(b)). Tab. 3 shows the quantitative comparison between the ordinary MVS method and our regularized one. Both qualitative and quantitative comparisons verify that this regularization eliminates noise and outliers and improves the quality of inference results.

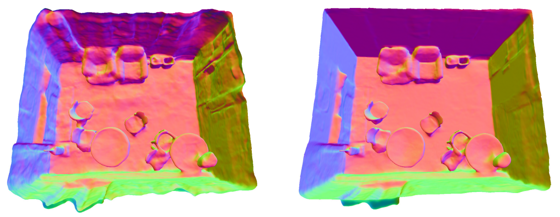
Analysis on the interval of learning iterations for intertwined regularization In the learning of HelixSurf, we set a hyperparameter N_{inter} as the interval of iterations for interwined regularization; HelixSurf performs MVS every

N^{th}_{inter} interation. Tab. 4 shows that N_{inter} less affects the performance of HelixSurf (and its convergence).

Efficacy of Sampling Guided by Dynamic Occupancy Grids To further alleviate the difficulties of optimization, we maintain dynamic occupancy grids and propose a sampling strategy to skip the sample points in empty space. The time consumption comparisons in Tab. 1 show that our training convergence is significantly faster than the existing neural implicit surface learning methods. The results in Tab. 5 present the timing consumptions of each part in the entire training process with and without the sampling strategy, respectively.

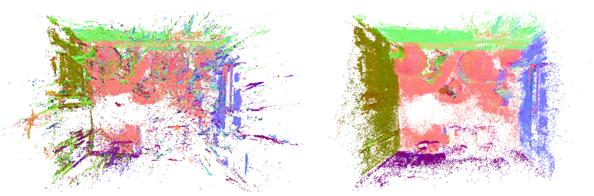
Regularization							
ordinary MVS	regularized MVS	Textureless Areas Handling	Acc↓	Comp↓	Prec↑	Recall↑	F-score↑
			0.179	0.208	0.313	0.275	0.291
✓			0.059	0.076	0.661	0.605	0.632
	✓		0.051	0.066	0.711	0.649	0.679
✓		✓	0.047	0.053	0.768	0.706	0.735
	✓	✓	**0.038**	**0.044**	**0.786**	**0.727**	**0.755**

Table 2. Analyses on the regularization of neural implicit surface learning from MVS predictions.



(a) W/O smoothness on textureless surface areas
(b) With smoothness on textureless surface areas

Figure 5. Visualization of example reconstruction without the use of smoothness scheme for textureless areas (a) and with the use of smoothness scheme (b). The colors encode surface normals.



(a) ordinary MVS.
(b) regularized MVS.

Figure 6. Qualitative comparisons of inference results between the ordinary MVS method (a) and our regularized MVS (b).

5.3. Real-world Large-scale Scene Reconstruction

To further examine the applicability and generalization of HelixSurf, we conduct experiments on an indoor subset from the Tanks and Temples [20] dataset. Results in Fig. 7(a, b) show that HelixSurf achieves reasonable results on such large-scale indoor scenes. Furthermore, we evaluate HelixSurf on large-scale outdoor scenes from Tanks and

Method	Depth map			
	Abs Diff↓	Abs Rel↓	Sq Rel↓	RMSE ↓
ordinary	0.067	0.098	0.020	0.147
regularized	0.053	0.085	0.011	0.106
Method	Normal map			
	Mean ↓	Median↓	RMSE↓	Prop_30° ↑
ordinary	35.5°	30.4°	42.6°	51.0%
regularized	27.8°	20.2°	35.3°	67.4%

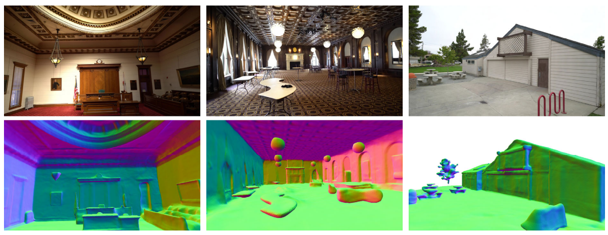
Table 3. Quantitative comparison between the ordinary MVS and our regularized MVS.

N_{inter}	4000	6000	8000	10000	12000
F-score↑	0.753	0.752	0.755	0.754	0.752

Table 4. Results with different intervals of learning iterations for intertwined regularization.

Occ Grids	MVS	Texture-less	Grid	Training Forward	Training Backward	**Total**
w/	3.8	2.6	0.6	12.4	13.8	**33.2**
w/o			-	184	203	393.4

Table 5. Time consumption (in minutes) for each part of our training process with or without guidance of Dynamic Occupancy Grids.



(a) (b) (c)

Figure 7. **Qualitative result of the reconstruction on Tanks and Temples [20].** (a) and (b) are examples of indoor scenes. (c) is an outdoor scene.

Temples [20]. Surprisingly, HelixSurf has potential to handle large-scale outdoor scenes, as shown in Fig. 7(c). Please refer to the supplementary for more results.

5.4. Conclusion

In this paper, we introduce an efficient and high-quality indoor scene reconstruction method, named HelixSurf. A novel intertwined regularization is proposed that benefits from both the traditional MVS and the neural implicit surface learning. We also propose a superpixel-based, adaptive scheme that regularizes the learning on textureless areas. Our design again confirms that combining traditional pipelines with the recent, differentiable rendering based neural learning can be helpful for surface reconstruction.

References

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012. 5

[2] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24, 2009. 3

[3] Michael Bleyer, Christoph Rhemann, and Carsten Rother. Patchmatch stereo-stereo matching with slanted support windows. In *Bmvc*, volume 11, pages 1–11, 2011. 3

[4] Kang Chen, Yu-Kun Lai, and Shi-Min Hu. 3d indoor scene modeling from rgb-d data: a survey. *Computational Visual Media*, 1(4):267–278, 2015. 1

[5] Julian Chibane, Gerard Pons-Moll, et al. Neural unsigned distance fields for implicit function learning. *Advances in Neural Information Processing Systems*, 33:21638–21652, 2020. 2, 3

[6] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 2, 6, 7

[7] Angela Dai, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM Transactions on Graphics (ToG)*, 36(4):1, 2017. 6

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

[9] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59(2):167–181, 2004. 5

[10] 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. 2

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

[12] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 873–881, 2015. 1, 2, 3

[13] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *Proceedings of the 37th International Conference on Machine Learning*, pages 3789–3799, 2020. 6

[14] 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. 2, 6

[15] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision.* Cambridge university press, 2003. 1, 2

[16] Sunghoon Im, Hae-Gon Jeon, Stephen Lin, and In So Kweon. Dpsnet: End-to-end deep plane sweep stereo. In *International Conference on Learning Representations*, 2018. 2

[17] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, volume 7, 2006. 1, 2, 4, 5

[18] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3):1–13, 2013. 1, 2, 4, 5

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[20] 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. 2, 6, 8

[21] Patrick Labatut, Jean-Philippe Pons, and Renaud Keriven. Efficient multi-view reconstruction of large-scale scenes using interest points, delaunay triangulation and graph cuts. In *2007 IEEE 11th international conference on computer vision*, pages 1–8. IEEE, 2007. 2

[22] Jiabao Lei and Kui Jia. Analytic marching: An analytic meshing solution from deep implicit surface networks. In *International Conference on Machine Learning*, pages 5789–5798. PMLR, 2020. 4

[23] Ruilong Li, Matthew Tancik, and Angjoo Kanazawa. Nerfacc: A general nerf acceleration toolbox. *arXiv preprint arXiv:2210.04847*, 2022. 6

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

[25] 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. 6

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

[27] 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. 2, 3

[28] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, pages 405–421, 2020. 1, 2, 3, 6

[29] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, July 2022. 6

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

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

[32] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 2, 3

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6

[34] 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. 2

[35] Andrea Romanoni and Matteo Matteucci. Tapa-mvs: Textureless-aware patchmatch multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10413–10422, 2019. 2

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

[37] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European conference on computer vision*, pages 501–518. Springer, 2016. 1, 2, 3, 5, 6

[38] Shuhan Shen. Accurate multiple view 3d reconstruction using patch-based stereo for large-scale scenes. *IEEE transactions on image processing*, 22(5):1901–1914, 2013. 2, 3

[39] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. Patchmatchnet: Learned multi-view patchmatch stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14194–14203, 2021. 2

[40] Jiepeng Wang, Peng Wang, Xiaoxiao Long, Christian Theobalt, Taku Komura, Lingjie Liu, and Wenping Wang. Neuris: Neural reconstruction of indoor scenes using normal priors. *arXiv preprint arXiv:2206.13597*, 2022. 2, 6

[41] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *Advances in Neural Information Processing Systems*, 34:27171–27183, 2021. 1, 2, 3, 4, 6

[42] Yiqun Wang, Ivan Skorokhodov, and Peter Wonka. Hf-neus: Improved surface reconstruction using high-frequency details. *arXiv preprint arXiv:2206.07850*, 2022. 2

[43] Qingshan Xu and Wenbing Tao. Multi-scale geometric consistency guided multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 5

[44] Qingshan Xu and Wenbing Tao. Planar prior assisted patchmatch multi-view stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12516–12523, 2020. 1, 2, 3, 6

[45] Qingshan Xu and Wenbing Tao. Pvsnet: Pixelwise visibility-aware multi-view stereo network. *arXiv preprint arXiv:2007.07714*, 2020. 2

[46] 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. 2

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

[48] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021. 1, 2, 3, 4, 6

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

[50] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2, 6

[51] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 2

[52] Enliang Zheng, Enrique Dunn, Vladimir Jojic, and Jan-Michael Frahm. Patchmatch based joint view selection and depthmap estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1510–1517, 2014. 1, 2, 3