Neural 3D Scene Reconstruction with the Manhattan-world Assumption

Haoyu Guo[1*] Sida Peng[1*] Haotong Lin[1] Qianqian Wang[2]
Guofeng Zhang[1] Hujun Bao[1] Xiaowei Zhou[1†]
[1] Zhejiang University [2] Cornell University

Abstract

This paper addresses the challenge of reconstructing 3D indoor scenes from multi-view images. Many previous works have shown impressive reconstruction results on textured objects, but they still have difficulty in handling low-textured planar regions, which are common in indoor scenes. An approach to solving this issue is to incorporate planer constraints into the depth map estimation in multi-view stereo-based methods, but the per-view plane estimation and depth optimization lack both efficiency and multi-view consistency. In this work, we show that the planar constraints can be conveniently integrated into the recent implicit neural representation-based reconstruction methods. Specifically, we use an MLP network to represent the signed distance function as the scene geometry. Based on the Manhattan-world assumption, planar constraints are employed to regularize the geometry in floor and wall regions predicted by a 2D semantic segmentation network. To resolve the inaccurate segmentation, we encode the semantics of 3D points with another MLP and design a novel loss that jointly optimizes the scene geometry and semantics in 3D space. Experiments on ScanNet and 7-Scenes datasets show that the proposed method outperforms previous methods by a large margin on 3D reconstruction quality. The code and supplementary materials are available at https://zju3dv.github.io/manhattan_sdf.

1. Introduction

Reconstructing 3D scenes from multi-view images is a cornerstone of many applications such as augmented reality, robotics, and autonomous driving. Given input images, traditional methods [43, 44, 58] generally estimate the depth map for each image based on the multi-view stereo (MVS) algorithms and then fuse estimated depth maps into 3D models. Although these methods achieve successful reconstruction in most cases, they have difficulty in handling low-textured regions, e.g., floors and walls of indoor scenes, due to the unreliable stereo matching in these regions.

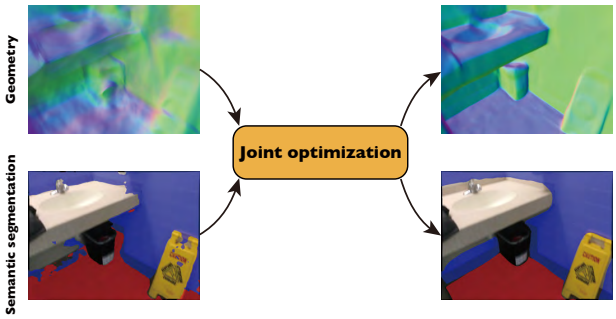

Figure 1. **Core idea.** We represent the geometry and semantics of 3D scenes with implicit neural representations, which enables the joint optimization of geometry reconstruction and semantic segmentation in 3D space based on the Manhattan-world assumption.

To improve the reconstruction of low-textured regions, a typical approach is leveraging the planar prior of man-made scenes, which has long been explored in literature [8, 10, 11, 41, 48, 51]. A renowned example is the Manhattan-world assumption [8], i.e., the surfaces of man-made scenes should be aligned with three dominant directions. These works either use plane estimation as a postprocessing step to inpaint the missing depth values in low-textured regions, or integrate planar constraints in stereo matching or depth optimization. However, all of them focus on optimizing per-view depth maps instead of the full scene models in 3D space. As a result, depth estimation and plane segmentation could still be inconsistent among views, yielding suboptimal reconstruction quality as demonstrated by our experimental results in Section 5.3.

There is a recent trend to represent 3D scenes as implicit neural representations [32, 46, 55] and learn the representations from images with differentiable renderers. In particular, [49, 54, 55] use a signed distance field (SDF) to represent the scene and render it into images based on the sphere tracing or volume rendering. Thanks to the well-defined surfaces of SDFs, they recover high-quality 3D ge-

The authors from Zhejiang University are affiliated with the State Key Lab of CAD&CG and the ZJU-SenseTime Joint Lab of 3D Vision. *Equal contribution. †Corresponding author: Xiaowei Zhou.

ometries from images. However, these methods essentially rely on the multi-view photometric consistency to learn the SDFs. So they still suffer from poor performance in low-textured planar regions, as shown in Figure 1, as many plausible solutions may satisfy the photometric constraint in low-textured planar regions.

In this work, we show that the Manhattan-world assumption [8] can be conveniently integrated into the learning of implicit neural representations of 3D indoor scenes and significantly improves the reconstruction quality. Unlike previous MVS methods that perform per-view depth optimization, implicit neural representations allow the joint representation and optimization of scene geometry and semantics simultaneously in 3D space, yielding globally-consistent reconstruction and segmentation. Specifically, we use an MLP network to predict signed distance, color and semantic logits for any point in 3D space. The semantic logits indicate the probability of a point being floor, wall or background, initialized by a 2D semantic segmentation network [4]. Similar to [54], we learn the signed distance and color fields by comparing rendered images to input images based on volume rendering. For the surface points on floors and walls, we enforce their surface normals to respect the Manhattan-world assumption. Considering the initial segmentation could be inaccurate, we design a loss that simultaneously optimizes the semantic logits along with the SDF. This loss effectively improves both the scene reconstruction and semantic segmentation, as illustrated in Figure 1.

We evaluate our method on the ScanNet [9] and 7-Scenes [45] datasets, which are widely-used datasets for 3D indoor scene reconstruction. The experiments show that the proposed approach outperforms the state-of-the-art methods in terms of reconstruction quality by a large margin, especially in planar regions. Furthermore, the joint optimization of semantics and reconstruction improves the initial semantic segmentation accuracy.

In summary, our contributions are as follows:

- A novel scene reconstruction approach that integrates the Manhattan-world constraint into the optimization of implicit neural representations.
- A novel loss function that optimizes semantic labels along with scene geometry.
- Significant gains of reconstruction quality compared to state-of-the-art methods on ScanNet and 7-Scenes.

2. Related work

MVS. Many methods adopt a two-stage pipeline for multi-view 3D reconstruction: first estimating the depth map for each image based on MVS and then performing depth fusion [27,31] to obtain the final reconstruction results. Traditional MVS methods [43,44] are able to reconstruct very accurate 3D shapes and have been used in many downstream applications such as novel view synthesis [39, 40]. However, they tend to give poor performance on texture-less regions. A major reason is that texture-less regions make dense feature matching intractable. To overcome this problem, some works improve the reconstruction pipeline with deep learning techniques. For instance, [15,52,53] attempt to extract image features, build cost volumes and use 3D CNNs to predict depth maps. [7, 13] construct cost volumes in a coarse-to-fine manner and can achieve high resolution results. Another line of works [11, 41, 48, 51] utilize scene priors to help the reconstruction. They observe that texture-less planar regions could be completed using planar prior. [19, 25, 56] propose a depth-normal consistency loss to improve training process. Instead of predicting the depth map for each image, our method learns an implicit neural representation, which can achieve more coherent and accurate reconstruction.

Neural scene reconstruction. Neural scene reconstruction methods predict the properties of points in the 3D space using neural networks. Atlas [30] presents an end-to-end reconstruction pipeline which directly regresses truncated signed distance function from the 3D feature volume. NeuralRecon [47] improves the reconstruction speed through reconstructing local surfaces for each fragment sequence. They represent scenes as discrete voxels, resulting in the high memory consumption. Recently, some methods [28, 29, 33, 34, 46, 49, 54] represent scenes with implicit neural functions and are able to produce high-resolution reconstruction with low memory consumption. [22, 32] propose an implicit differentiable renderer, which enables learning 3D shapes from 2D images. IDR [55] models view-dependent appearance and can be applied to non-Lambertian surface reconstruction. Despite achieving impressive performance, they need mask information to obtain the reconstruction. Inspired by the success of NeRF [29], NeuS [49] and VolSDF [54] attach volume rendering techniques to IDR and eliminate the need for mask information. Although they achieve amazing reconstruction results of scenes with small scale and rich textures, we experimentally find that these methods tend to produce poor results in large scale indoor scenes with texture-less planar regions. In contrast, our method utilizes semantic information to assist reconstruction in texture-less planar regions.

Semantic segmentation. Recently, learning-based methods have achieved impressive progress on semantic segmentation. FCN [24] applies fully convolution on the whole image to produce pixel-level image semantic segmentation results. Recent methods [2, 6] attempt to aggregate high-resolution feature maps using a learnable decoder to keep the detailed spatial information in the deep layers. Another line of works [4, 5, 57] use dilated convolutions for large re-

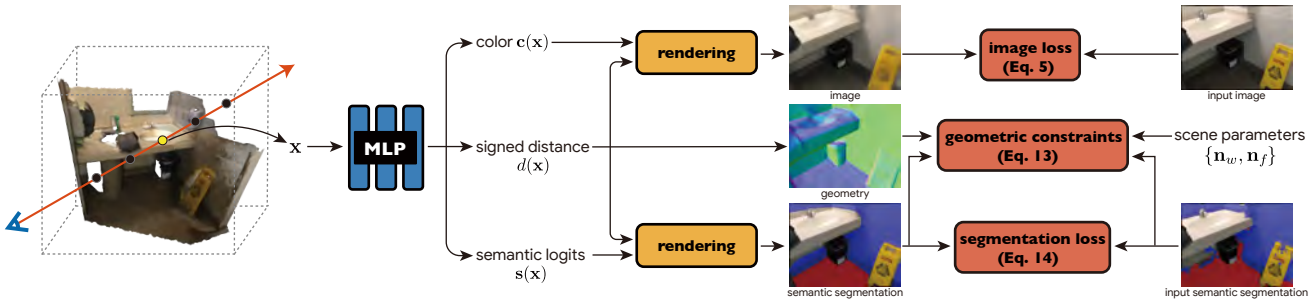

Figure 2. **Overview of our method.** We learn the geometry, appearance and semantics of 3D scenes with implicit neural representations. For an image pixel, we use differentiable volume rendering to render its pixel color and semantic probabilities, which are supervised with input images and semantic labels in 2D. To jointly optimize the geometry and semantics, we introduce geometric constraints in planar regions based on the Manhattan-world assumption, which improves both the reconstruction and segmentation accuracy.

ceptive fields. In addition to 2D segmentation methods, a lot of works aim to achieve semantic segmentation from 3D space. [3,36–38] develop networks to process different representations of 3D data including point clouds and voxels. More recently, [59] proposes to extend NeRF to encode semantics with radiance fields. The intrinsic multi-view consistency and smoothness of NeRF benefit semantics, which enables label propogation, super-resolution, denoising and several tasks. There are also some works [14, 18, 20, 23] that learn semantic segmentation in both 2D and 3D space and utilize the projection relation between images and 3D scenes to facilitate the performance. Our method learns 3D semantics from 2D segmentation prediction [5] and jointly optimizes semantics with geometry.

3. Method

Given multi-view images with camera poses of an indoor scene, our goal is to reconstruct the high-quality scene geometry. In this paper, we propose a novel approach called ManhattanSDF, as illustrated in Figure 2. We represent the scene geometry and appearance with signed distance and color fields, which are learned from images with volume rendering techniques (Sec. 3.1). To improve the reconstruction quality in texture-less regions (e.g., walls and floors), we perform semantic segmentation to detect these regions and apply the geometric constraints based on the Manhattan-world assumption [8] (Sec. 3.2). To overcome the inaccuracy of semantic segmentation, we additionally encode the semantic information into the implicit scene representation and jointly optimize the semantics together with the geometry and appearance of the scene (Sec. 3.3).

3.1. Learning scene representations from images

In contrast to MVS methods [44,52], we model the scene as an implicit neural representation and learn it from images with a differentiable renderer. Inspired by [49, 54, 55], we represent the scene geometry and appearance with signed distance and color fields. Specifically, given a 3D point $\mathbf{x}$, the geometry model maps it to a signed distance $d(\mathbf{x})$, which is defined as:

$$(d(\mathbf{x}), \mathbf{z}(\mathbf{x})) = F_d(\mathbf{x}), \tag{1}$$

where F_d is implemented as an MLP network, and $\mathbf{z}(\mathbf{x})$ is the geometry feature as in [55]. To approximate the radiance function, the appearance model takes the spatial point $\mathbf{x}$, the view direction $\mathbf{v}$, the normal $\mathbf{n}(\mathbf{x})$, and the geometry feature $\mathbf{z}(\mathbf{x})$ as inputs and outputs color $\mathbf{c}(\mathbf{x})$, which is defined as:

$$\mathbf{c}(\mathbf{x}) = F_{\mathbf{c}}(\mathbf{x}, \mathbf{v}, \mathbf{n}(\mathbf{x}), \mathbf{z}(\mathbf{x})), \tag{2}$$

where we obtain the normal $\mathbf{n}(\mathbf{x})$ by computing the gradient of the signed distance $d(\mathbf{x})$ at point $\mathbf{x}$ as in [55].

Following [49, 54], we adopt volume rendering to learn the scene representation networks from images. Specifically, to render an image pixel, we sample N points $\{\mathbf{x}_i\}$ along its camera ray $\mathbf{r}$. Then we predict the signed distance and color for each point. To apply volume rendering techniques, we transform the signed distance $d(\mathbf{x})$ to the volume density $\sigma(\mathbf{x})$:

$$\sigma(\mathbf{x}) = \begin{cases} \frac{1}{\beta}\left(1 - \frac{1}{2}\exp\left(\frac{d(\mathbf{x})}{\beta}\right)\right) & \text{if } d(\mathbf{x}) < 0, \\ \frac{1}{2\beta}\exp\left(-\frac{d(\mathbf{x})}{\beta}\right) & \text{if } d(\mathbf{x}) \geq 0, \end{cases} \tag{3}$$

where β is a learnable parameter. Then we accumulate the densities and colors using numerical quadrature [29]:

$$\hat{\mathbf{C}}(\mathbf{r}) = \sum_{i=1}^{K} T_i (1 - \exp(-\sigma_i \delta_i)) \mathbf{c}_i, \tag{4}$$

where $\delta_i = ||\mathbf{x}_{i+1} - \mathbf{x}_i||_2$ is the distance between adjacent sampled points, and $T_i = \exp(-\sum_{j=1}^{i-1} \sigma_j \delta_j)$ denotes the accumulated transmittance along the ray.

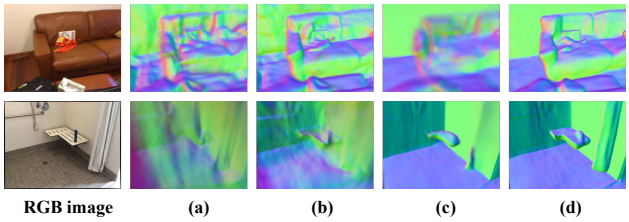

Figure 3. **Qualitative ablations.** (a) Training with only images. (b) Adding $\mathcal{L}_d$. (c) Adding $\mathcal{L}_{\text{geo}}$. (d) Replacing $\mathcal{L}_{\text{geo}}$ with $\mathcal{L}_{\text{joint}}$.

During training, we optimize the scene representation networks using multi-view images with photometric loss:

$$\mathcal{L}_{\text{img}} = \sum_{\mathbf{r}\in\mathcal{R}} \left\| \hat{\mathbf{C}}(\mathbf{r}) - \mathbf{C}(\mathbf{r}) \right\|, \tag{5}$$

where $\mathbf{C}(\mathbf{r})$ is the ground-truth pixel color, and $\mathcal{R}$ is the set of camera rays going through sampled pixels. Additionally, we apply Eikonal loss [12] as suggested by [54, 55].

$$\mathcal{L}_E = \sum_{\mathbf{y}\in\mathcal{Y}} (\|\nabla_{\mathbf{y}} d(\mathbf{y})\|_2 - 1)^2, \tag{6}$$

where $\mathcal{Y}$ denotes the combination of points sampled from random uniform space and surface points for pixels.

We observe that learning the scene representation from scratch with only images has difficulty in reconstructing reasonable geometries even in textured regions, as shown in Figure 3(a). In contrast, although depth estimation based methods [43,44,58] tend to give incomplete reconstructions in low-textured regions, they can reconstruct accurate point clouds of textured regions from images. We propose to use depth maps from multi-view stereo method [43] to assist the learning of the scene representations:

$$\mathcal{L}_d = \sum_{\mathbf{r}\in\mathcal{D}} \left| \hat{D}(\mathbf{r}) - D(\mathbf{r}) \right|, \tag{7}$$

where $\mathcal{D}$ is the set of camera rays going through image pixels that have depth values estimated by [43], $\hat{D}(\mathbf{r})$ and $D(\mathbf{r})$ are rendered and input depth values, respectively. Figure 3(b) presents an example of the reconstruction result using the depth loss. Although the depth loss improves the reconstruction quality, the reconstruction performance is still limited in texture-less regions, since input depth maps are incomplete in these regions.

3.2. Scene reconstruction with planar constraints

We observe that most texture-less planar regions lie on floors and walls. As pointed by the Manhattan-world assumption [8], floors and walls of indoor scenes generally align with three dominant directions. Motivated by this, we propose to apply the geometric constraints to the regions of floors and walls. Specifically, we first use a 2D semantic segmentation network [5] to obtain the regions of floors and walls. Then we apply loss functions to enforce the surface points in a planar region to share the same normal direction.

For the supervision of floor regions, we assume that floors are vertical to the z-axis following the Manhattan-world assumption. We design the normal loss for a floor pixel as:

$$\mathcal{L}_f(\mathbf{r}) = \left| 1 - \mathbf{n}(\mathbf{x}_\mathbf{r}) \cdot \mathbf{n}_f \right|, \tag{8}$$

where $\mathbf{x}_\mathbf{r}$ is the surface intersection point of camera ray $\mathbf{r}$, $\mathbf{n}(\mathbf{x}_\mathbf{r})$ is the normal calculated as the gradient of signed distance $d(\mathbf{x})$ at point $\mathbf{x}_\mathbf{r}$, and $\mathbf{n}_f = \langle 0, 0, 1 \rangle$ is an upper unit vector that denotes the assumed normal direction in the floor regions.

To supervise the wall regions, a learnable normal $\mathbf{n}_w$ is introduced. We design a loss that enforces the normal directions of surface points on walls to be either parallel or orthogonal with the learnable normal $\mathbf{n}_w$, which is defined for wall pixels as:

$$\mathcal{L}_w(\mathbf{r}) = \min_{i\in\{-1,0,1\}} \left| i - \mathbf{n}(\mathbf{x}_\mathbf{r}) \cdot \mathbf{n}_w \right|, \tag{9}$$

where the learnable normal $\mathbf{n}_w$ is initialized as $\langle 1, 0, 0 \rangle$ and is jointly optimized with network parameters during training. We fix the last element of $\mathbf{n}_w$ as 0 to force it vertical to $\mathbf{n}_f$. Finally, we define the normal loss as:

$$\mathcal{L}_{\text{geo}} = \sum_{r\in\mathcal{F}} \mathcal{L}_f(\mathbf{r}) + \sum_{r\in\mathcal{W}} \mathcal{L}_w(\mathbf{r}), \tag{10}$$

where $\mathcal{F}$ and $\mathcal{W}$ are the sets of camera rays of image pixels that are predicted as floor and wall regions by the semantic segmentation network [5].

3.3. Joint optimization of semantics and geometry

Applying geometric constraints to floor and wall regions improves the reconstruction quality. However, 2D semantic segmentation results predicted by the network could be wrong in some image regions, which leads to inaccurate reconstruction, as shown in Figure 3(c). To solve this problem, we propose to optimize semantic labels in 3D together with scene geometry and appearance.

Inspired by [59], we augment the neural scene representation by additionally predicting semantic logits for each point in 3D space. Let us denote semantic logits for $\mathbf{x}$ as $\mathbf{s}(\mathbf{x}) \in \mathbb{R}^3$. The semantic logits are defined as:

$$\mathbf{s}(\mathbf{x}) = F_\mathbf{s}(\mathbf{x}), \tag{11}$$

where $F_\mathbf{s}$ is an MLP network. By applying softmax function, the logits can be transformed to the probabilities of point $\mathbf{x}$ being floor, wall and other regions. Similar to image rendering, we render the semantic logits into 2D image space with volume rendering techniques. For an image

pixel, its semantic logits are obtained by:

$$\hat{\mathbf{S}}(\mathbf{r}) = \sum_{i=1}^{N} T_i(1 - \exp(-\sigma_i \delta_i))\mathbf{s}_i, \tag{12}$$

where $\mathbf{s}_i$ is the logits of sampled point $\mathbf{x}_i$ along the camera ray $\mathbf{r}$. We forward the logits $\hat{\mathbf{S}}$ into a softmax normalization layer to compute the multi-class probabilities $\hat{p}_f$, $\hat{p}_w$ and $\hat{p}_b$, denoting the probabilities of the pixel being floor, wall and other regions.

During training, we integrate the multi-class probabilities into the geometric losses proposed in Sec. 3.2. To this end, we improve the normal loss in Equation (10) to a joint optimization loss, which is defined as:

$$\mathcal{L}_{\text{joint}} = \sum_{\mathbf{r}\in\mathcal{F}} \hat{p}_f(\mathbf{r})\mathcal{L}_f(\mathbf{r}) + \sum_{\mathbf{r}\in\mathcal{W}} \hat{p}_w(\mathbf{r})\mathcal{L}_w(\mathbf{r}). \tag{13}$$

This loss function optimizes the scene representation in the following way. Taking the floor region as an example, if the input semantic label of $\mathbf{r}$ is correct, $\mathcal{L}_f(\mathbf{r})$ should decrease easily. But if the input segmentation is wrong, $\mathcal{L}_f(\mathbf{r})$ could vibrate during training. To decrease $\hat{p}_f(\mathbf{r})\mathcal{L}_f(\mathbf{r})$, the gradient will push $\hat{p}_f(\mathbf{r})$ to be small, which thus optimizes the semantic label. Note that a trivial solution is that both $\hat{p}_f$ and $\hat{p}_w$ vanish. To avoid this, we also supervise the semantics with input semantic segmentation results estimated by [5] using the cross entropy loss:

$$\mathcal{L}_{\mathbf{s}} = -\sum_{\mathbf{r}\in\mathcal{R}} \sum_{k\in\{f,w,b\}} p_k(\mathbf{r}) \log \hat{p}_k(\mathbf{r}), \tag{14}$$

where $\hat{p}_k(\mathbf{r})$ is the rendered probability for class k and $p_k(\mathbf{r})$ is 2D semantic segmentation prediction. Note that learning 3D semantics with $\mathcal{L}_{\mathbf{s}}$ naturally utilizes the multi-view consensus to improve the accuracy of semantic scene segmentation, as shown in [59].

4. Implementation details

We implement our method with PyTorch [35] and adopt DeepLabV3+ [5] from Detectron2 [50] for implementing 2D semantic segmentation network. The network training is performed on one NVIDIA TITAN Xp GPU. In experiments, we first normalize all cameras to be inside a unit sphere and initialize network parameters following [1] so that the SDF is approximated to a uint sphere, and the surface normals of the sphere are facing inside. Images are resized to 640×480 for both 2D semantic segmentation and scene reconstruction. We use Adam optimizer [17] with learning rate of 5e-4 and train the network with batches of 1024 rays for 50k iterations. The optimization can be completed in about 5 hours for each scene. We use Marching Cubes algorithm [26] for extracting surface mesh from the learned signed distance function.

5. Experiments

5.1. Datasets, metrics and baselines

Datasets. We perform the experiments on ScanNet (V2) [9] and 7-Scenes [45]. ScanNet is an RGB-D video dataset that contains 1613 indoor scenes with 2.5 million views. It is annotated with ground-truth camera poses, surface reconstructions, and instance-level semantic segmentations. 7-Scenes consists of RGB-D frames recorded by a hand-held Kinect RGB-D camera. It uses KinectFusion to obtain camera poses and dense 3D models. In our experiments, we train the 2D semantic segmentation network on training set of ScanNet and perform the experiments on 8 randomly selected scenes (4 from validation set of ScanNet and 4 from 7-Scenes). Each scene contains 1K–5K views. We uniformly sample one tenth views for reconstruction.

Metrics. For 3D reconstruction, we use RGB-D fusion results as ground truth and evaluate our method using 5 standard metrics defined in [30]: accuracy, completeness, precision, recall and F-score. We consider F-score as the overall metric following [47]. The definitions of these metrics are detailed in the supplementary material. For semantic segmentation, we evaluate Intersection over Union (IoU) of floor and wall.

Baselines. (1) Classical MVS method: COLMAP [43]. We use screened Poisson Surface reconstruction (sPSR) [16] to reconstruct mesh from point clouds. (2) MVS methods with plane fitting: COLMAP*. There are several methods [11, 41] that segment piece-wise plane segmentations in image space and apply plane fitting to COLMAP. Since these methods have not released code, we implement this baseline using state-of-the-art piece-wise plane segmentation method [21] and denote it as COLMAP*. (3) MVS method with plane regularization: ACMP [51]. ACMP utilizes a probabilistic graphical model to embed planar models into PatchMatch and proposes multi-view aggregated matching cost to improve depth estimation in planar regions. (4) State-of-the-art volume rendering based methods: NeRF [29], UNISURF [33], NeuS [49] and VolSDF [54]. For these methods, we use Marching Cubes algorithm [26] to extract mesh. Since they (including our method) can reconstruct unobserved regions which will be penalized in evaluation, we render depth maps from predicted mesh and re-fuse them using TSDF fusion [31] following [30].

5.2. Ablation studies

We conduct ablation studies on ScanNet and show the effectiveness of each component in our method.

We train with four configurations: (1) raw setting of **VolSDF**: training network with only image supervision, (2) **VolSDF-D**: we add depth supervision $\mathcal{L}_d$ defined in Section 3.1, (3) **VolSDF-D-G**: in addition to VolSDF-D, we add

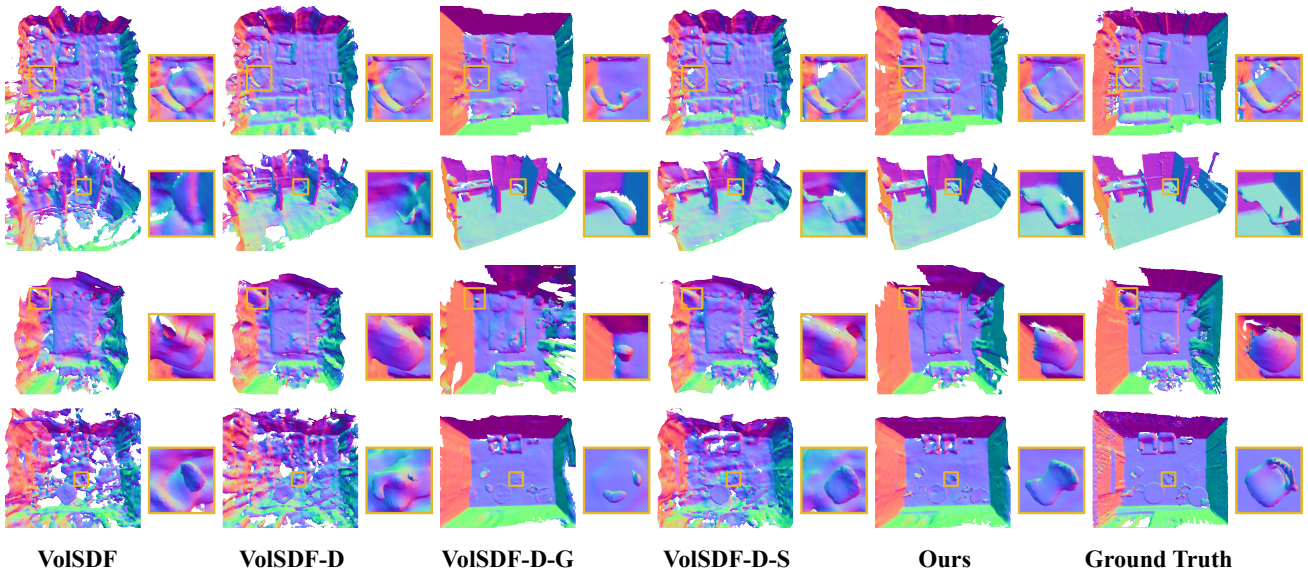

Figure 4. **Ablation studies on ScanNet.** Our method can produce much more coherent reconstruction results compared to our baselines. Note that **VolSDF-D-G** can reconstruct smoother and more complete planes compared to **VolSDF** and **VolSDF-D**. **Ours** can maintain the reconstruction quality of planes while also reconstruct much more details in non-planar regions compared to **VolSDF-D-G**. The color indicates surface normal. Zoom in for details.

	Acc↓	Comp↓	Prec↑	Recall↑	F-score↑
VolSDF	0.414	0.120	0.321	0.394	0.346
VolSDF-D	0.229	0.099	0.416	0.455	0.431
VolSDF-D-G	0.133	0.090	0.447	0.435	0.438
VolSDF-D-S	0.127	0.081	0.463	0.487	0.474
Ours	**0.072**	**0.068**	**0.621**	**0.586**	**0.602**

Table 1. **Ablation studies on ScanNet.** We report 3D reconstruction metrics. Our method has a notable improvement in terms of both accuracy and completeness compared to our baselines.

normal loss $\mathcal{L}_{\text{geo}}$ defined in Section 3.2, (4) **VolSDF-D-S**: in addition to VolSDF-D, we learn semantics in 3D space, (5) **Ours**: we learn semantics in 3D space and improve normal loss to joint optimization loss $\mathcal{L}_{\text{joint}}$ defined in Section 3.3. We report quantitative results in Table 1 and provide qualitative results in Figure 4.

Comparing **VolSDF** and **VolSDF-D** in Table 1 shows that supervision from estimated sparse depth maps gives about 0.095 precision improvement and 0.061 recall improvement. Visualization results in Figure 4 show that there are improvements in both planar and non-planar regions, but the reconstruction is still noisy and incomplete. These results demonstrate that $\mathcal{L}_d$ can make network converge better but the reconstruction results are still of low quality.

Then, we study how the normal loss affects the reconstruction performance. Results in Table 1 show that **VolSDF-D-G** gives 0.031 precision improvement, but recall is decreased by 0.020. As shown in visualization results in Figure 4, **VolSDF-D-G** can reconstruct smoother and more complete planes compared to **VolSDF-D**, but some details of non-planar regions are missed. These results demonstrate that $\mathcal{L}_{\text{geo}}$ can improve the reconstruction in planar regions, but the performance in non-planar regions could be decreased due to misleading of wrong segmentation.

To validate the benefit of learning semantic fileds, we compare **VolSDF-D** and **VolSDF-D-S**. Results in Table 1 show that **VolSDF-D-S** gives 0.047 precision improvement and 0.032 recall improvement. These results demonstrate that learning semantics in 3D space can also assist reconstruction.

To validate the benefit of our proposed joint optimization manner, we compare **VolSDF-D-G** and **Ours** in Table 4. Substituting $\mathcal{L}_{\text{geo}}$ with $\mathcal{L}_{\text{joint}}$ gives 0.174 precision improvement and 0.151 recall improvement. Visualization results in Figure 4 show that, while **Ours** can keep great reconstruction performance in planar regions, the reconstruction in non-planar regions are also improved notably. These results demonstrate that **Ours** can achieve the most coherent reconstruction results.

5.3. Comparisons with the state-of-the-art methods

3D reconstruction. We evaluate 3D geometry metrics on ScanNet and 7-Scenes. Averaged quantitative results are shown in Table 2. Please refer to the supplementary material for detailed results on individual scenes. Qualitative results on ScanNet are shown in Figure 5. By analysing quantitative and qualitative results, we found that our method

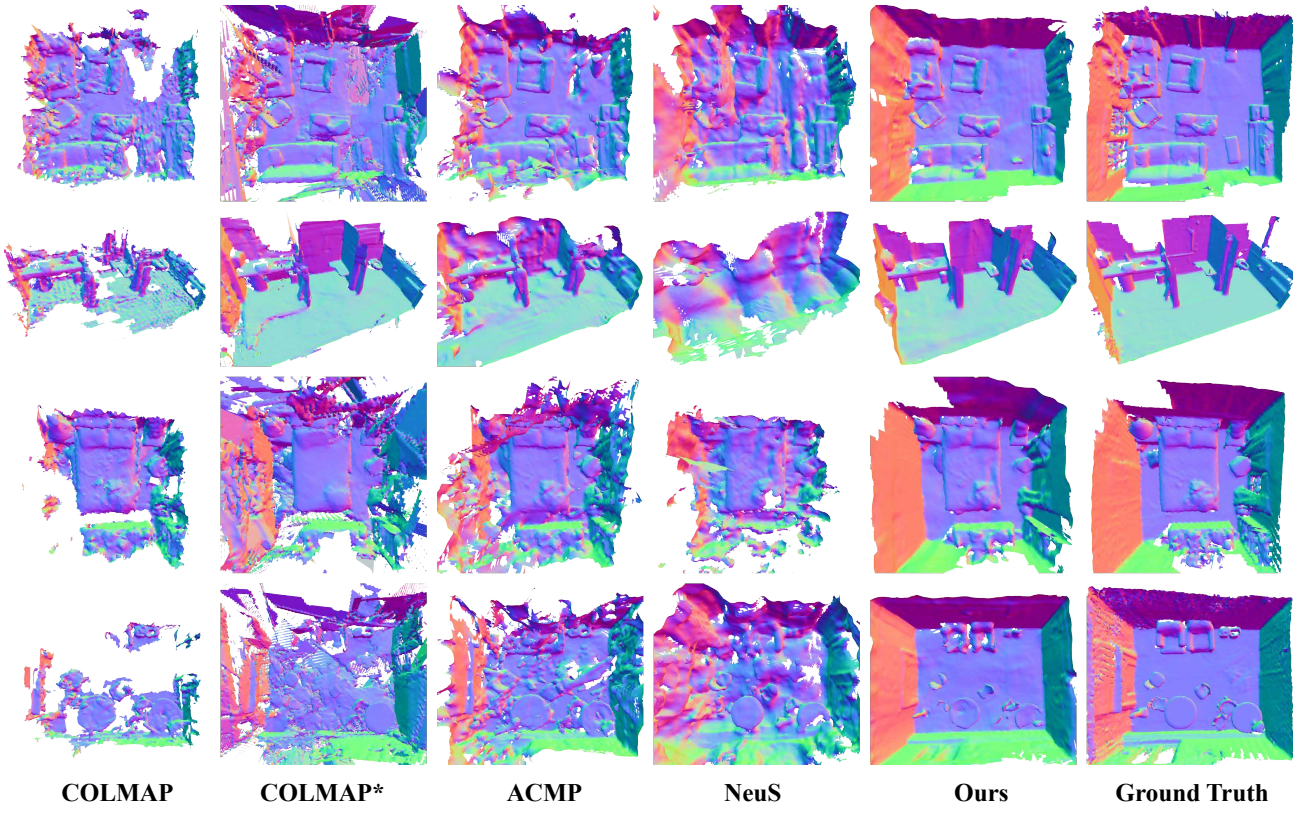

Figure 5. **3D reconstruction results on ScanNet.** Our method significantly outperforms COLMAP and volume rendering-based methods. Furthermore, compared with methods that apply planar prior to MVS, we can produce more coherent reconstruction results especially in planar regions. Zoom in for details.

Method	ScanNet					7-Scenes				
	Acc↓	Comp↓	Prec↑	Recall↑	**F-score↑**	Acc↓	Comp↓	Prec↑	Recall↑	**F-score↑**
COLMAP	**0.047**	0.235	**0.711**	0.441	0.537	**0.069**	0.417	**0.536**	0.202	0.289
COLMAP*	0.396	0.081	0.271	**0.595**	0.368	0.670	0.215	0.116	0.215	0.149
ACMP	0.118	0.081	0.531	0.581	0.555	0.293	0.194	0.350	0.269	0.299
NeRF	0.735	0.177	0.131	0.290	0.176	0.573	0.321	0.159	0.085	0.083
UNISURF	0.554	0.164	0.212	0.362	0.267	0.407	0.136	0.195	0.301	0.231
NeuS	0.179	0.208	0.313	0.275	0.291	0.151	0.247	0.313	0.229	0.262
VolSDF	0.414	0.120	0.321	0.394	0.346	0.285	0.140	0.220	0.285	0.246
Ours	0.072	**0.068**	0.621	0.586	**0.602**	0.112	**0.133**	0.351	**0.326**	**0.336**

Table 2. **Averaged 3D reconstruction metrics on ScanNet and 7-Scenes.** We compare our method with MVS and volume rendering based methods. The accuracy of our method ranks only second to COLMAP and our completeness is on par with MVS methods with planar prior. Considering both accuracy and completeness, our method achieves the best reconstruction performance.

significantly outperforms state-of-the-art MVS and volume rendering based methods considering both reconstruction precision and recall.

COLMAP can achieve extremely high precision as it filters out reconstructed points which are inconsistent between multiple views in fusion stage. However, this process sacrifices recall. COLMAP* and ACMP can obviously complete some missing areas and acheive higher recall by applying planar prior to COLMAP. However, their optimization strategy can not guarantee the consistency of estimated depth maps, resulting in noisy reconstructions. The performance of NeRF is poor since the volume density representation has not sufficient constraint on geometry. Other volume rendering based methods – UNISURF, NeuS and VolSDF perform better than NeRF as occupancy and signed distance function have better surface constraints. However, they still struggle

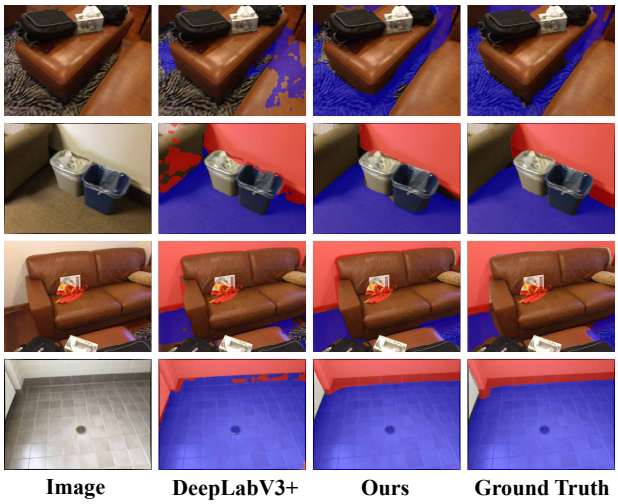

Figure 6. **Semantic segmentation results.** We compare our semantic segmentation results with DeepLabV3+. We mask pixels of floor and wall labels with blue and red.

	IoU^f ↑	IoU^w ↑	IoU^m ↑
DeepLabV3+	0.532	0.475	0.503
Ours	**0.624**	**0.518**	**0.571**

Table 3. **Quantitative results of semantic segmentation.** IoU^f and IoU^w denote IoU of floor and wall regions, respectively. IoU^m denotes the average of IoU^f and IoU^w.

in reconstructing accurate and complete geometry.

Semantic segmentation. We render our learned semantics to image space and evaluate semantic segmentation metrics on ScanNet. We compare our method with DeepLabV3+. and report quantitative results in Table 3. Qualitative results are shown in Figure 6. Quantitatively, our metrics in both floor and wall regions are improved distinctly compared to DeepLabV3+. Visualization results show that semantic segmentation results predicted by DeepLabV3+ have non-negligible noise especially near boundaries. The noise are ruleless and generally inconsistent between different views. By learning semantics in 3D space, our method can naturally combine multi-view information and improve consistency so that the noise could be remitted notably. However, there are also some misclassified pixels that cannot be easily corrected using multi-view consistency. Taking the last row of Figure 6 for example, the bottom of wall has different color with the main part of the wall, so that some pixels are wrongly recognized as floor. This kind of phenomenon can occur in every view and is different from the inconsistent noise. By optimizing semantics together with geometry, these misclassified pixels could be corrected.

Novel view synthesis. Our accurate reconstruction results enable us to render high-quality images under novel views. To evaluate novel view synthesis, we select some novel views away from training views and render images. The qualitative comparison are shown in Figure 7. More results can be found in the supplementary material.

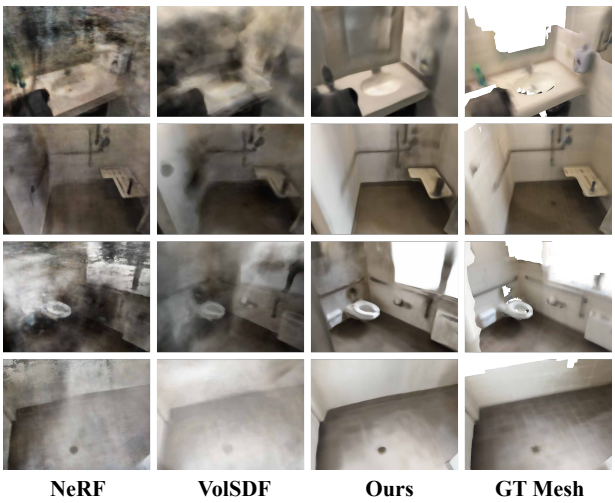

Figure 7. **Novel view synthesis results.** We select novel views relatively far from training views for the qualitative comparison. Our method produces better rendering results compared to NeRF and VolSDF. Due to the lack of ground truth images in novel views, we render GT mesh in these views for reference.

6. Conclusion

In this paper, we introduced a novel indoor scene reconstruction method based on the Manhattan-world assumption. The key idea is to utilize semantic information in planar regions to guide geometry reconstruction. Our method learns 3D semantics from 2D segmentation results, and jointly optimizes 3D semantics with geometry to improve the robustness against inaccurate 2D segmentation. Experiments showed that the proposed method was able to reconstruct accurate and complete planes while maintaining details of non-planar regions, and significantly outperformed the state-of-the-art methods on public datasets.

Limitations. This work only considers the Manhattan-world assumption. While most man-made scenes obey this assumption, some scenarios require a more general assumption, e.g., the Atlanta-world assumption [42]. The proposed framework could be extended to adopt other assumptions by modifying the formulation of geometric constraints in the loss function.

Acknowledgement. The authors would like to acknowledge the support from the National Key Research and Development Program of China (No. 2020AAA0108901), NSFC (No. 62172364), and the ZJU-SenseTime Joint Lab of 3D Vision.

References

[1] Matan Atzmon and Yaron Lipman. SAL: Sign agnostic learning of shapes from raw data. In *CVPR*, 2020. 5

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *T-PAMI*, 2017. 2

[3] Alexandre Boulch, Bertrand Le Saux, and Nicolas Audebert. Unstructured point cloud semantic labeling using deep segmentation networks. *3DOR*, 2017. 3

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *T-PAMI*, 2017. 2

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 2, 3, 4, 5

[6] Bowen Cheng, Liang-Chieh Chen, Yunchao Wei, Yukun Zhu, Zilong Huang, Jinjun Xiong, Thomas S Huang, Wen-Mei Hwu, and Honghui Shi. SPGNet: Semantic prediction guidance for scene parsing. In *ICCV*, 2019. 2

[7] Shuo Cheng, Zexiang Xu, Shilin Zhu, Zhuwen Li, Li Erran Li, Ravi Ramamoorthi, and Hao Su. Deep stereo using adaptive thin volume representation with uncertainty awareness. In *CVPR*, 2020. 2

[8] James M Coughlan and Alan L Yuille. Manhattan world: Compass direction from a single image by bayesian inference. In *ICCV*, 1999. 1, 2, 3, 4

[9] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *CVPR*, 2017. 2, 5

[10] Yasutaka Furukawa, Brian Curless, Steven M Seitz, and Richard Szeliski. Manhattan-world stereo. In *CVPR*, 2009. 1

[11] David Gallup, Jan-Michael Frahm, and Marc Pollefeys. Piecewise planar and non-planar stereo for urban scene reconstruction. In *CVPR*, 2010. 1, 2, 5

[12] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *T-PAMI*, 2020. 4

[13] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *CVPR*, 2020. 2

[14] Wenbo Hu, Hengshuang Zhao, Li Jiang, Jiaya Jia, and Tien-Tsin Wong. Bidirectional projection network for cross dimension scene understanding. In *CVPR*, 2021. 3

[15] Sunghoon Im, Hae-Gon Jeon, Stephen Lin, and In So Kweon. DPSNet: End-to-end deep plane sweep stereo. In *ICLR*, 2019. 2

[16] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM TOG*, 2013. 5

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[18] Abhijit Kundu, Xiaoqi Yin, Alireza Fathi, David Ross, Brian Brewington, Thomas Funkhouser, and Caroline Pantofaru. Virtual multi-view fusion for 3D semantic segmentation. In *ECCV*, 2020. 3

[19] Uday Kusupati, Shuo Cheng, Rui Chen, and Hao Su. Normal assisted stereo depth estimation. In *CVPR*, 2020. 2

[20] Felix Järemo Lawin, Martin Danelljan, Patrik Tosteberg, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Deep projective 3D semantic segmentation. In *CAIP*, 2017. 3

[21] Chen Liu, Kihwan Kim, Jinwei Gu, Yasutaka Furukawa, and Jan Kautz. PlaneRCNN: 3D plane detection and reconstruction from a single image. In *CVPR*, 2019. 5

[22] 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 *CVPR*, 2020. 2

[23] Zhengzhe Liu, Xiaojuan Qi, and Chi-Wing Fu. 3D-to-2D distillation for indoor scene parsing. In *CVPR*, 2021. 3

[24] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 2

[25] Xiaoxiao Long, Lingjie Liu, Christian Theobalt, and Wenping Wang. Occlusion-aware depth estimation with adaptive normal constraints. In *ECCV*, 2020. 2

[26] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3D surface construction algorithm. *SIGGRAPH*, 1987. 5

[27] Paul Merrell, Amir Akbarzadeh, Liang Wang, Philippos Mordohai, Jan-Michael Frahm, Ruigang Yang, David Nistér, and Marc Pollefeys. Real-time visibility-based fusion of depth maps. In *ICCV*, 2007. 2

[28] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3D reconstruction in function space. In *CVPR*, 2019. 2

[29] 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 *ECCV*, 2020. 2, 3, 5

[30] 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 *ECCV*, 2020. 2, 5

[31] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. KinectFusion: Real-time dense surface mapping and tracking. In *ISMAR*, 2011. 2, 5

[32] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3D representations without 3D supervision. In *CVPR*, 2020. 1, 2

[33] Michael Oechsle, Songyou Peng, and Andreas Geiger. UNISURF: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *ICCV*, 2021. 2, 5

[34] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019. 2

[35] 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. In *NeurIPS*, 2019. 5

[36] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *CVPR*, 2017. 3

[37] Charles R Qi, Hao Su, Matthias Nießner, Angela Dai, Mengyuan Yan, and Leonidas J Guibas. Volumetric and multi-view cnns for object classification on 3D data. In *CVPR*, 2016. 3

[38] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*, 2017. 3

[39] Gernot Riegler and Vladlen Koltun. Free view synthesis. In *ECCV*, 2020. 2

[40] Gernot Riegler and Vladlen Koltun. Stable view synthesis. In *CVPR*, 2021. 2

[41] Andrea Romanoni and Matteo Matteucci. TAPA-MVS: Textureless-aware patchmatch multi-view stereo. In *ICCV*, 2019. 1, 2, 5

[42] Grant Schindler and Frank Dellaert. Atlanta world: An expectation maximization framework for simultaneous low-level edge grouping and camera calibration in complex man-made environments. In *CVPR*, 2004. 8

[43] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 1, 2, 4, 5

[44] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 2016. 1, 2, 3, 4

[45] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *CVPR*, 2013. 2, 5

[46] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3D-structure-aware neural scene representations. In *NeurIPS*, 2019. 1, 2

[47] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. NeuralRecon: Real-time coherent 3D reconstruction from monocular video. In *CVPR*, 2021. 2, 5

[48] Shang Sun, Yunan Zheng, Xuelei Shi, Zhenyu Xu, and Yiguang Liu. PHI-MVS: Plane hypothesis inference multi-view stereo for large-scale scene reconstruction. *arXiv*, 2021. 1, 2

[49] 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. In *NeurIPS*, 2021. 1, 2, 3, 5

[50] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 5

[51] Qingshan Xu and Wenbing Tao. Planar prior assisted patchmatch multi-view stereo. In *AAAI*, 2020. 1, 2, 5

[52] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. MVSNet: Depth inference for unstructured multi-view stereo. In *ECCV*, 2018. 2, 3

[53] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *CVPR*, 2019. 2

[54] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *NeurIPS*, 2021. 1, 2, 3, 4, 5

[55] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Ronen Basri, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In *NeurIPS*, 2020. 1, 2, 3, 4

[56] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *ICCV*, 2019. 2

[57] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 2

[58] Enliang Zheng, Enrique Dunn, Vladimir Jojic, and Jan-Michael Frahm. Patchmatch based joint view selection and depthmap estimation. In *CVPR*, 2014. 1, 4

[59] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *ICCV*, 2021. 3, 4, 5