Neural 3D Scene Reconstruction With Indoor Planar Priors

Xiaowei Zhou, Haoyu Guo, Sida Peng, Yuxi Xiao, Haotong Lin, Qianqian Wang, Guofeng Zhang, and Hujun Bao

***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 planar 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 and the Atlanta-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.**

***Index Terms*—3D reconstruction, implicit neural representations, the manhattan-world assumption, the Atlanta-world assumption.**

I. 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 [56], [57], [77] 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.

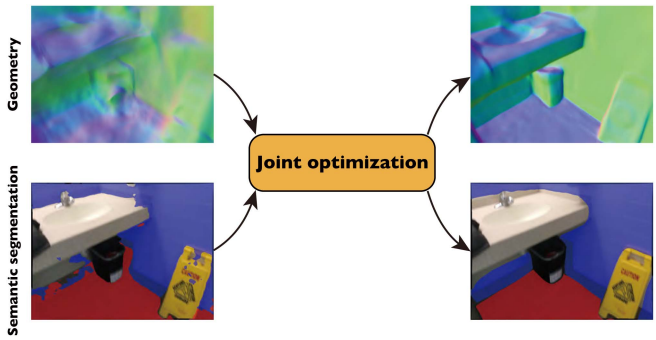


Fig. 1. Main idea. We represent the geometry and semantics of 3D scenes with implicit neural representations, which enable the joint optimization of geometry reconstruction and semantic segmentation in 3D space based on indoor planar assumptions.

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 [10], [15], [16], [54], [62], [68]. A renowned example is the Manhattan-world assumption [10], 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 V-C.

There is a recent trend to represent 3D scenes as implicit neural representations [43], [59], [72] and learn the representations from images with differentiable renderers. In particular, [65], [71], [72] 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 geometries 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 Fig. 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 [10] and the Atlanta-world assumption [55] can be

Manuscript received 18 June 2023; revised 25 January 2024; accepted 15 March 2024. Date of publication 20 March 2024; date of current version 6 August 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2020AAA0108901 and in part by the Information Technology Center and State Key Lab of CAD&CG, Zhejiang University. Recommended for acceptance by V. Lepetit. *(Corresponding author: Hujun Bao.)*

Xiaowei Zhou, Haoyu Guo, Sida Peng, Yuxi Xiao, Haotong Lin, Guofeng Zhang, and Hujun Bao are with Zhejiang University, Hangzhou 310027, China (e-mail: xwzhou@zju.edu.cn; guohaoyu@zju.edu.cn; pengsida@zju.edu.cn; yuxixiao@zju.edu.cn; haotongl@zju.edu.cn; zhangguofeng@zju.edu.cn; baohujun@zju.edu.cn).

Qianqian Wang is with UC Berkeley, Berkeley, CA 94720 USA (e-mail: qianqianwang@berkeley.edu).

The code and supplementary materials are available at https://zju3dv.github.io/manhattan_sdf.

Digital Object Identifier 10.1109/TPAMI.2024.3379833

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 [6]. Similar to [71], 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 indoor planar 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 Fig. 1.

We evaluate our method on the ScanNet [11] and 7-Scenes [58] 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 indoor scene priors 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.

A preliminary version of this work appeared in CVPR 2022 [20]. Here, the work is extended in the following ways. First, we extend our method to more general indoor scenes with the Atlanta-world assumption by proposed adaptive normal clustering strategy. Second, inspired by [40], we integrate spatial hash encoding into our method, which substantially accelerate the training process. Third, additional experiments on non-Manhattan scenes of ScanNet [11] are conducted to evaluate our approach, we also evaluate our extended method on Manhattan scenes to show backward compatibility of our extension. Moreover, we perform detailed ablation studies to validate the effectiveness of each extension.

II. 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 [36], [42] to obtain the final reconstruction results. Traditional MVS methods [56], [57] are able to reconstruct very accurate 3D shapes and have been used in many downstream applications such as novel view synthesis [52], [53]. 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, [23], [69], [70] attempt to extract image features, build cost volumes and use 3D CNNs to predict depth maps. [9], [19] construct cost volumes in a coarse-to-fine manner and can achieve high resolution results. Another line of works [16], [54], [62], [68] utilize scene priors to help the reconstruction. They observe that texture-less planar regions could be completed using planar prior. [27], [34], [73] 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 [41] presents an end-to-end reconstruction pipeline which directly regresses truncated signed distance function from the 3D feature volume. NeuralRecon [61] 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 [37], [38], [44], [45], [59], [65], [66], [71] represent scenes with implicit neural functions and are able to produce high-resolution reconstruction with low memory consumption. [31], [43] propose an implicit differentiable renderer, which enables learning 3D shapes from 2D images. IDR [72] 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 [38], NeuS [65] and VolSDF [71] attach volume rendering techniques to IDR and eliminate the need for mask information. Mip-NeRF [3] proposes a multiscale representation that addresses inherent aliasing to achieve better performance. MonoSDF [75] leverages learned monocular depth and normal prior to improve multi-view geometric reconstruction and achieve appealing results. However, its performance heavily depends on the quality of the learned monocular depth and normal estimation. [47] also exploits the Manhattan-world assumption for neural scene reconstruction. Our work and [47] share similar ideas, but we address quite different problems. Our work expands the Manhattan-SDF from the Manhattan-world assumption to the more general Atlanta-world assumption, whereas [47] is still based on the Manhattan-world assumption, extending the constraints to the areas beyond floors and walls, e.g., tables.

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.

Some methods attempt to accelerate NeRF in either rendering speed or the optimization process. NSVF [30] utilizes an explicit sparse voxel structure to skip empty voxels during ray marching and achieve efficient rendering. KiloNeRF [51] replaces the large MLP in NeRF with thousands of tiny MLPs.

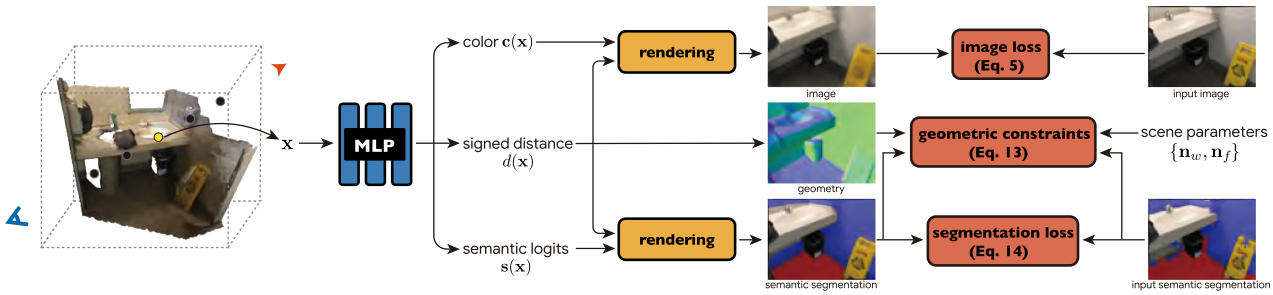


Fig. 2. Overview of our method. We learn the geometry, appearance and semantics of 3D scenes with implicit neural representations. For an image pixel, we use differentiable volume rendering to render its pixel color and semantic probabilities, which are supervised with input images and semantic labels 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.

FastNeRF [17] and SqueezeNeRF [64] factorize NeRF into several subnetworks and build a cached representation which can be efficiently queried during rendering. SNeRG [21] and PlenOctrees [74] convert NeRF to explicit volumetric representation for inference, which supports real-time rendering. While these methods can achieve high-speed rendering, they still require a long time for training. Plenoxels [14] and DVGO [60] directly adopt the explicit volumetric representation during training and utilize coarse to fine strategy to achieve high resolution. Instant-NGP [40] employs a multiresolution hash table for encoding and lightweight MLPs as decoder, achieving unprecedentedly fast training. TensoRF [5] and CCNeRF [63] utilize tensor decomposition to factorize the feature grid into compact components and lead to high memory efficiency, as well as fast training. K-Planes [13] further models temporal information, thereby enabling more effective reconstruction of dynamic scenes.

Semantic segmentation: Recently, learning-based methods have achieved impressive progress on semantic segmentation. FCN [33] applies fully convolution on the whole image to produce pixel-level image semantic segmentation results. Recent methods [2], [8] 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 [6], [7], [76] use dilated convolutions for large receptive fields. In addition to 2D segmentation methods, a lot of works aim to achieve semantic segmentation from 3D space. [4], [48], [49], [50] develop networks to process different representations of 3D data including point clouds and voxels. More recently, [78] proposes to extend NeRF to encode semantics with radiance fields. The intrinsic multi-view consistency and smoothness of NeRF benefit semantics, which enables label propagation, super-resolution, denoising and several tasks. There are also some works [22], [26], [28], [32] 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 [7] and jointly optimizes semantics with geometry.

III. 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 as illustrated in Fig. 2. We represent the scene geometry and appearance with signed distance and color fields, which are learned from images with volume rendering techniques (Section III-A). 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 indoor planar assumption [10], [55] (Section III-B). 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 (Section III-C).

A. Learning Scene Representations From Images

In contrast to MVS methods [57], [69], we model the scene as an implicit neural representation and learn it from images with a differentiable renderer. Inspired by [65], [71], [72], 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 [72]. 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 [72].

Following [65], [71], 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 [38]

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

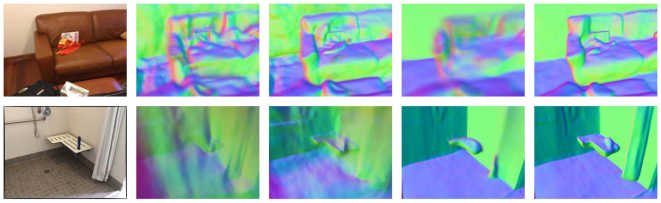


Fig. 3. Qualitative ablations. (a) Training with only images. (b) Adding $\mathcal{L}_d$. (c) Adding $\mathcal{L}_{\text{geo}}$. (d) Replacing $\mathcal{L}_{\text{geo}}$ with $\mathcal{L}_{\text{joint}}$.

TABLE I
ABLATION STUDIES ON SCANNET

	Acc↓	Comp↓	Prec↑	Recall↑	F-score↑
VolSDF	0.414	0.120	0.321	0.394	0.346
VolSDF-D	0.069	0.067	0.614	0.591	0.601
VolSDF-D-G	0.073	0.058	0.664	0.647	0.655
VolSDF-D-S	0.065	0.062	0.627	0.614	0.621
Ours	**0.053**	**0.056**	**0.714**	**0.664**	**0.688**

We report 3D reconstruction metrics. Our method has a notable improvement in terms of both accuracy and completeness compared to our baselines. Bold value means the best result among all methods.

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.

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 [18] as suggested by [71], [72].

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

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 Fig. 3(a). In contrast, although depth estimation based methods [56], [57], [77] 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 [56] 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 [56], $\hat{D}(\mathbf{r})$ and $D(\mathbf{r})$ are rendered and input depth values, respectively. Fig. 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.

B. Scene Reconstruction With Planar Constraints

We observe that most texture-less planar regions lie on floors and walls. While it is hard to reconstruct these regions directly, we found strong prior exists in man-made scenes, which can be formulated as geometric constraints on floors and walls. Specifically, we adopt The Manhattan-world assumption [10] and The Atlanta-world assumption [55], which have long been explored in literature, and design corresponding geometric constraints.

1) The Manhattan-World Assumption: As pointed by the Manhattan-world assumption [10], 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 [7] 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

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

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

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

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 [7].

2) The Atlanta-World Assumption: While most man-made scenes obey the Manhattan-world assumption, some scenarios require a more general assumption. To make our method more generalizable, we relax the constraints and adopt the Atlanta-world assumption [55]. Similar to Manhattan-world assumption, the Atlanta-world assumption [55] states that floors are vertical to z-axis and walls are aligned with z-axis. But it assumes that the relative angles among the walls are arbitrary. Due to the unknown number of those arbitrary walls' normal directions, it is difficult to define and jointly optimize the normal directions with the SDF function as done in Manhattan-world scenes in Section III-B1. Therefore, we propose a coarse-to-fine **adaptivenormalclusteringstrategy** to automatically learn the normal directions of walls.

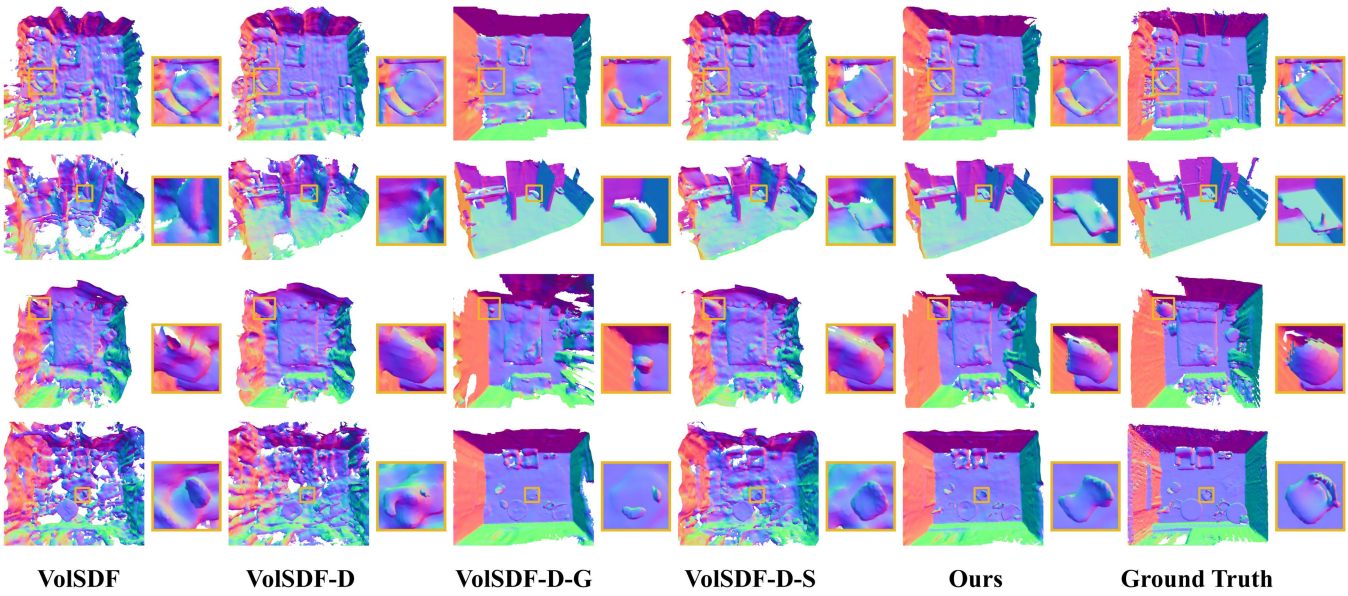


Fig. 4. Ablation studies on ScanNet. Our method can produce much more coherent reconstruction results compared to our baselines. Note that *VolSDF-D-G* can reconstruct smoother and more complete planes compared to *VolSDF* and *VolSDF-D*. *Ours* can maintain the reconstruction quality of planes and also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

TABLE II
ABLATION STUDIES OF ADAPTIVE NORMAL CLUSTERING AND OUR ACCELERATION MODULE ON SCANNET

Normal clustering	Acceleration	Manhattan-world scenes					Atlanta-world scenes					Training time (hrs)↓
		Acc↓	Comp↓	Prec↑	Recall↑	F-score↑	Acc↓	Comp↓	Prec↑	Recall↑	F-score↑	
✗	✗	0.053	0.056	0.714	0.664	0.688	0.110	0.115	0.484	0.434	0.458	5.0
✓	✗	**0.048**	**0.052**	**0.718**	**0.673**	**0.695**	0.107	**0.110**	**0.498**	**0.445**	**0.470**	5.0
✓	✓	0.064	0.067	0.644	0.612	0.627	**0.104**	0.164	0.460	0.377	0.414	**0.7**

We evaluate the methods on both manhattan-world scenes and atlanta-world scenes. The experiments show that the design of adaptive normal clustering can extend our method to atlanta-world scenes, which can also work well on manhattan-world scenes. Moreover, the acceleration module can achieve similar reconstruction quality with a much faster training speed. Bold value means the best result among all methods.

TABLE III
ABLATION STUDIES OF DIFFERENT TYPES OF DEPTH SUPERVISION

	Acc↓	Comp↓	Prec↑	Recall↑	F-score↑
COLMAP depth	0.053	0.056	0.714	0.664	0.688
OmniData depth	0.043	0.050	0.743	0.688	0.714

We report the reconstruction results by using depth from COLMAP and depth predicted from monocular images by OmniData as our supervison.

Because of the unknown number of wall directions in the Atlanta-world assumption, we first uniformly initialize a set of learnable normal directions $\mathcal{N}_w = \{\mathbf{n}_{w1}, \mathbf{n}_{w2}, \mathbf{n}_{w3}, \ldots\}$. According to the Atlanta-world assumption, the wall normal loss is reformulated as

$$\mathcal{L}_w = \min_{\mathbf{n}_{wj} \in \mathcal{N}_w} \min_{i \in \{-1,0,1\}} |i - \mathbf{n}(\mathbf{x_r}) \cdot \mathbf{n}_{wj}|, \qquad (11)$$

where the normal direction of a wall point $\mathbf{n}(\mathbf{x_r})$ will query the closest learnable normal direction in $\mathcal{N}_w$ to calculate normal loss. We optimize the neural network parameters and the learnable normal directions in $\mathcal{N}_w$ together during training.

To ensure the reconstructed walls to be flat, $\mathcal{N}_w$ needs to be compact. To this end, we record the number of times each normal direction in $\mathcal{N}_w$ being queried during the calculation of $\mathcal{L}_w$, and periodically discard normal directions queried infrequently. For the remaining normal directions, we also conduct a periodic merge operation which merges two normal directions if their distance is smaller than a threshold. Please refer to Section IV for more details.

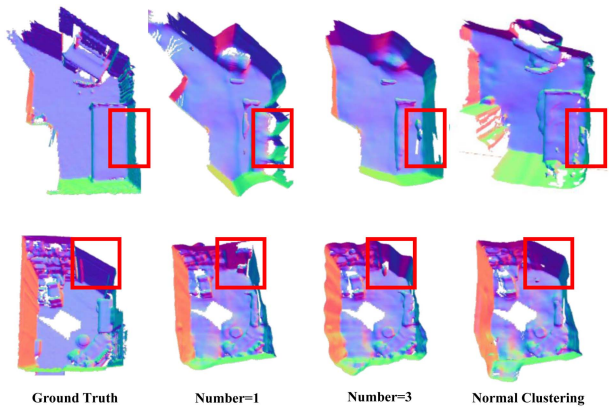


Fig. 5. Ablation studies for adaptive normal clustering strategy on ScanNet. We conduct the ablation studies for adaptive normal clustering strategy on Atlanta-world scenes. Specifically, we reduce the number of learnable normal directions to 1 and 3 and train the model without adaptive normal clustering strategy for comparison.

C. Joint Optimization of Semantics and Geometry

Applying geometric constraints to floor and wall regions improves the reconstruction quality. However, 2D semantic

TABLE IV
AVERAGED 3D RECONSTRUCTION METRICS ON MANHATTAN SCENES ON SCANNET AND 7-SCENES

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	0.053	**0.056**	**0.714**	**0.664**	**0.688**	0.112	**0.133**	0.351	**0.326**	**0.336**

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. Bold value means the best result among all methods.

TABLE V
AVERAGED 3D RECONSTRUCTION METRICS OF NON-MANHATTAN SCENES ON SCANNET

	Acc↓	Comp↓	Prec↑	Recall↑	F-score↑
COLMAP	0.083	0.276	0.501	0.335	0.393
COLMAP*	0.486	**0.161**	0.170	0.376	0.234
ACMP	**0.096**	0.196	**0.540**	0.381	0.439
NeRF	0.390	0.370	0.166	0.164	0.154
UNISURF	0.359	0.320	0.213	0.197	0.190
NeuS	0.192	0.433	0.255	0.175	0.206
VolSDF	0.259	0.347	0.252	0.200	0.219
Ours	0.101	0.180	0.475	**0.410**	**0.440**

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. Bold value means the best result among all methods.

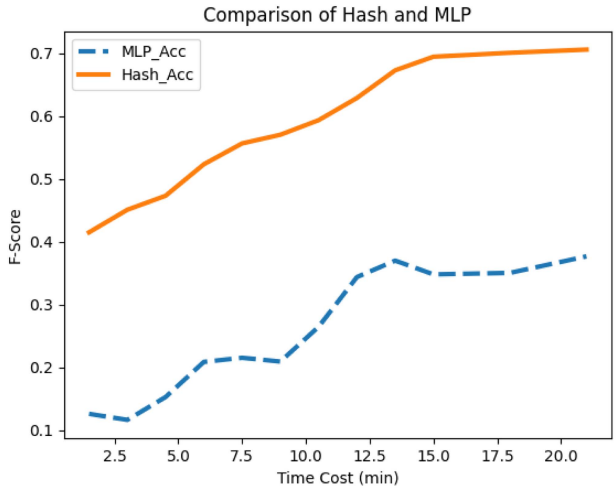


Fig. 6. Ablation studies for two architectures on ScanNet. Hash encoding-based architecture accelerates the convergence speed, which achieves convergence after 20 minutes of training with 20,000 iterations, while the MLP-based architecture trains much slower.

segmentation results predicted by the network could be wrong in some image regions, which leads to inaccurate reconstruction, as shown in Fig. 3(c). To solve this problem, we propose to optimize semantic labels in 3D together with scene geometry and appearance.

Inspired by [78], 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

$$\mathbf{s}(\mathbf{x}) = F_{\mathbf{s}}(\mathbf{x}), \tag{12}$$

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

$$\hat{\mathbf{S}}(\mathbf{r}) = \sum_{i=1}^{N} T_i(1 - \exp(-\sigma_i \delta_i))\mathbf{s}_i, \tag{13}$$

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 Section III-B. To this end, we improve the normal loss in (10) to a joint optimization loss, which is defined as

$$\mathcal{L}_{\text{joint}} = \sum_{\mathbf{r}\in\mathcal{F}} \hat{p}_f(\mathbf{r})\mathcal{L}_f(\mathbf{r}) + \sum_{\mathbf{r}\in\mathcal{W}} \hat{p}_w(\mathbf{r})\mathcal{L}_w(\mathbf{r}). \tag{14}$$

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 [7] using the cross entropy loss

$$\mathcal{L}_{\mathbf{s}} = -\sum_{\mathbf{r}\in\mathcal{R}} \sum_{k\in\{f,w,b\}} p_k(\mathbf{r}) \log \hat{p}_k(\mathbf{r}), \tag{15}$$

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 [78].

Accelerating the optimization process: Similar to [65], [71], [72], our method requires expensive time for optimization. To

make our method more applicable, we adopt the multi-resolution hash table [40] to reduce the training time. We construct multiresolution grids of L levels and a hash table consisting of T code. For each node $\mathbf{y}$ of the grids, we use a spatial hash function $h : \mathbb{Z}^d \mapsto \mathbb{Z}_T$ to index it to the hash table. The hash function is defined as

$$h(\mathbf{y}) = \left(\bigoplus_{i=1}^{3} y_i \pi_i \right) \mod T, \tag{16}$$

where $\oplus$ denotes the bit-wise XOR operation and π_i are unique, large prime numbers. After that, the encoding of an arbitrary input coordinate $\mathbf{x}$ at level l can be calculated by linear interpolation from gird points using weight $\mathbf{w}_l = \mathbf{x}_l - \lfloor \mathbf{x}_l \rfloor$. Lastly, we concatenate the feature vector of $\mathbf{x}$ at each level, as well as the encoded view direction $\mathbf{v}$, and feed them into a tiny MLP to predict the values. The hash encoding and feature interpolation step can be performed efficiently, and the tiny MLP costs much less time than the original MLP, so we can significantly accelerate the optimization process. Please refer to [40] for more details.

IV. Implementation Details

We implement our method with PyTorch [46]. The network training is performed on one NVIDIA TITAN Xp GPU. Images are resized to 640×480 for both 2D semantic segmentation and scene reconstruction. We take **VolSDF** [71] as our backbone, and our basic version is constructed with a simple Multi-layer perceptron neural network(MLP). Meanwhile, we implement our acceleration version based on multi-resolution hash table [39].

Initialization: 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 unit sphere, and we modify the sign to make sure that the surface normal of the sphere is facing inside.

Adaptive normal clustering strategy: We initialize 20 learnable normal directions at the beginning of the training, which evenly lie in the range of $[0, \frac{1}{2}\pi]$ (only needing a quarter part of search space, which is similar to (9)). During training, we record the number of queries made to each normal direction in $\mathcal{N}_w$ and determine the query frequency for each by dividing its query count by the sum of query counts of all normal directions. To implement periodic discard and merge, for every 50 iterations, we sort all normal directions in descending order of query frequency, select a subset of directions with the highest frequencies such that their cumulative frequency accounts for 0.9 and discard the others, then merge two normal directions if their L1 distance is smaller than 0.055. We set these hyperparameters empirically and find that our method is quite robust to them.

Multiresolution hash encoding: We set the number of levels of the multiresolution grids as $l = 8$, the resolution of the coarsest level as $N_l = 16$, the growth factor as $b = 1.38$, and the size of the hash table as $T = 2^{17}$. We use $\pi_1 = 1, \pi_2 = 2654435761, \pi_3 = 805459861$ in the spatial hash function.

Training strategy: We use Adam optimizer [25] with learning rate of 5e-4 to train the network with batches of 1,024 rays. Our approach requires 5-hour optimization (50 k iterations) for each scene without acceleration, which can be substantially reduced to 40 minutes (20 k iterations) by employing the acceleration module. After training, we use Marching Cubes algorithm [35] for extracting surface mesh from the learned signed distance function.

Semantic segmentation: We adopt DeepLabV3+ [7] from Detectron2 [67] for implementing 2D semantic segmentation network. We train the network on the training set of ScanNet [11] according to the official train/val/test split.

TABLE VI
Quantitative Results of Semantic Segmentation

	$\text{IoU}^f \uparrow$	$\text{IoU}^w \uparrow$	$\text{IoU}^m \uparrow$
DeepLabV3+	0.532	0.475	0.503
Ours	**0.624**	**0.518**	**0.571**

IoU^f and IoU^w denote IoU of floor and wall regions, respectively. IoU^m denotes the average of IoU^f and IoU^w. Bold value means the best result among all methods.

V. Experiments

A. Datasets, Metrics and Baselines

Datasets: We perform the experiments on ScanNet (V2) [11] and 7-Scenes [58]. 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 handheld 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 1 K–5 K 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 [41]: accuracy, completeness, precision, recall and F-score. We consider F-score as the overall metric following [61]. The definitions of these metrics are detailed in the supplementary material on our project page. For semantic segmentation, we evaluate Intersection over Union (IoU) of floor and wall.

Baselines: (1) Classical MVS method: COLMAP [56]. We use screened Poisson Surface reconstruction (sPSR) [24] to reconstruct mesh from point clouds. (2) MVS methods with plane fitting: COLMAP*. There are several methods [16], [54] 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 [29] and denote it as COLMAP*. (3) MVS method with plane regularization: ACMP [68]. 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 [38], UNISURF [44], NeuS [65] and VolSDF [71]. For these methods, we use Marching Cubes algorithm [35] to extract mesh. Since they (including our method) can reconstruct

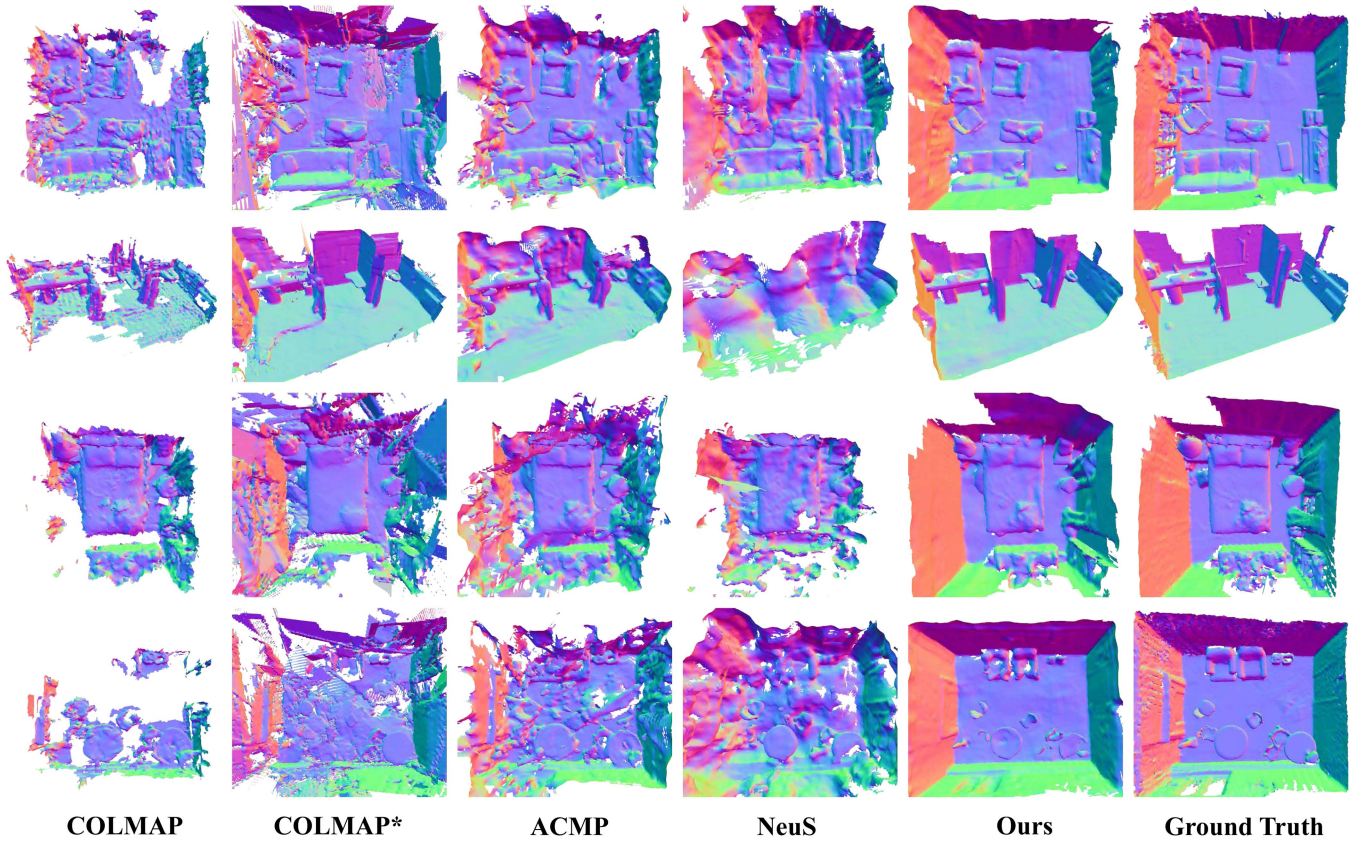


Fig. 7. 3D reconstruction results on Manhattan scenes on ScanNet. Our method significantly outperforms COLMAP and volume rendering-based methods. Furthermore, compared with methods that apply planar prior to MVS, we can produce more coherent reconstruction results especially in planar regions. Zoom in for details.

unobserved regions which will be penalized in evaluation, we render depth maps from predicted mesh and re-fuse them using TSDF fusion [42] following [41].

B. Ablation Studies

We conduct the ablation studies on ScanNet. To evaluate the effectiveness of our proposed geometric constraints based on the Manhattan-world assumption as well as joint optimization strategy, we train our basic model 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 III-A, (3) *VolSDF-D-G*: in addition to VolSDF-D, we add normal loss $\mathcal{L}_{\text{geo}}$ defined in Section III-B, (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 III-C. We report quantitative results in Table I and provide qualitative results in Fig. 4.

Comparing *VolSDF* and *VolSDF-D* in Table I shows that supervision from estimated sparse depth maps gives about 0.293 precision improvement and 0.197 recall improvement. Visualization results in Fig. 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 much better, but the reconstruction results are still of low quality.

Then, we study how the normal loss affects the reconstruction performance. Results in Table I show that *VolSDF-D-G* gives 0.050 precision improvement and 0.056 recall improvement. As shown in visualization results in Fig. 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 fields, we compare *VolSDF-D* and *VolSDF-D-S*. Results in Table I show that *VolSDF-D-S* gives 0.013 precision improvement and 0.023 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 I. Substituting $\mathcal{L}_{\text{geo}}$ with $\mathcal{L}_{\text{joint}}$ gives 0.050 precision improvement and 0.017 recall improvement. Visualization results in Fig. 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.

We conducted an additional ablation experiment to show the impact of different types of depth supervision on reconstruction results. The results in Table III indicate that using the depth predicted by OmniData [12] indeed improves the outcomes, but the enhancement is not significant.

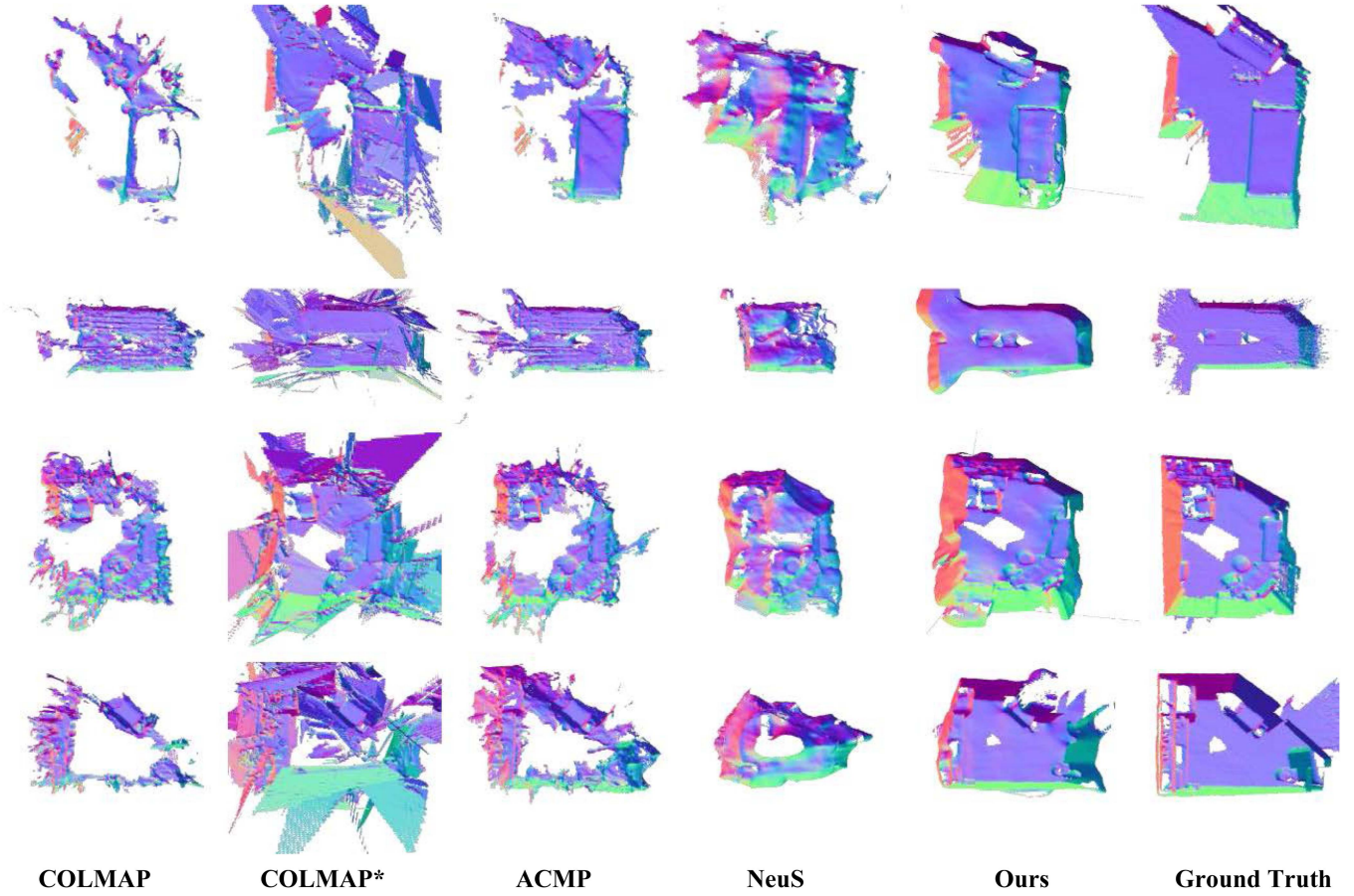


Fig. 8. 3D reconstruction results on non-Manhattan scenes on ScanNet. Our method significantly outperforms COLMAP and volume rendering-based methods. Furthermore, compared with methods that apply planar prior to MVS, we can produce more coherent reconstruction results especially in planar regions. Zoom in for details.

We also perform ablation studies on the proposed adaptive normal clustering strategy in both Manhattan-world and Atlanta-world scenes. Results in Fig. 5 qualitatively show that the learned normal directions can automatically fit the ground truth normal directions of the walls and efficiently improve the reconstruction quality of the Atlanta-world scenes. For comparison, we reduce the number of learnable normal directions to 1 (Manhattan-assumption) and 3 and train the model without adaptive normal clustering strategy, we found that they will suffer degraded performance. When $N = 1$, the decline in reconstruction quality is primarily manifested in walls that do not conform to the Manhattan-world assumption. These walls are poorly reconstructed, often appearing jagged, while other regions are almost unaffected. Therefore, the overall decrease in reconstruction quality is not particularly significant. Results in Table II quantitatively justify the effectiveness of our adaptive normal clustering strategy in our experiments. It is shown that the Atlanta-world assumption is more general and is also applicable to the Manhattan scenes (as special cases of Atlanta scenes). For non-Manhattan scenes, our Atlanta-based approach outperforms the Manhattan-based approach. For Manhattan scenes, the Atlanta-based approach performs comparably, and in some cases even slightly better than the Manhattan-based approach. Our observation is that many Manhattan scenes do not strictly satisfy the orthogonal constraint, and the Atlanta-world assumption provides more flexibility, allowing for more accurate learning of the normal directions of walls.

To show the effectiveness of the multi-resolution hash table, we report the convergence curve in Fig. 6 and the final geometric results in Table II. The results show that the hash-based model impressively shorten the training time while sacrificing little accuracy of reconstruction.

C. Comparisons With the State-of-the-Art Methods

3D reconstruction: For Manhattan-world scenes, we evaluate 3D geometry metrics on ScanNet and 7-Scenes. For Atlanta-world scenes, we only evaluate on ScanNet due to there are no Atlanta-world scene in 7-Scenes dataset. Averaged quantitative results are shown in Tables IV and V. Please refer to the supplementary material on our project page for detailed results on individual scenes. Qualitative results on ScanNet are shown in Figs. 7 and 8. By analyzing quantitative and qualitative results, we found that our method 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

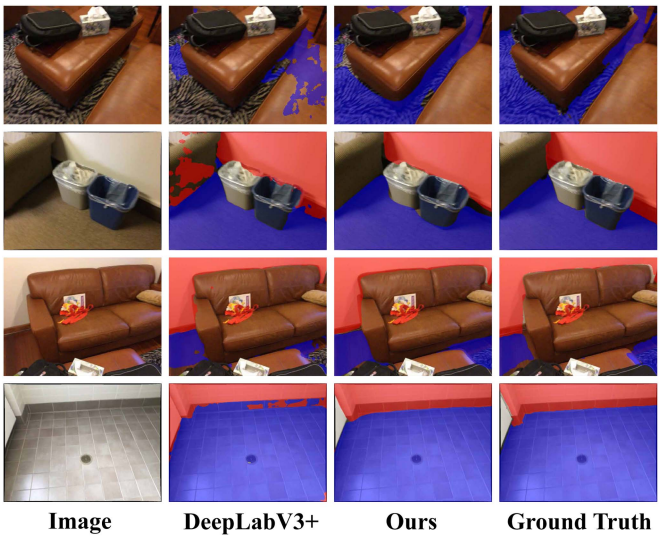


Fig. 9. Semantic segmentation results. We compare our semantic segmentation results with DeepLabV3+. We mask pixels of floor and wall labels with blue and red.

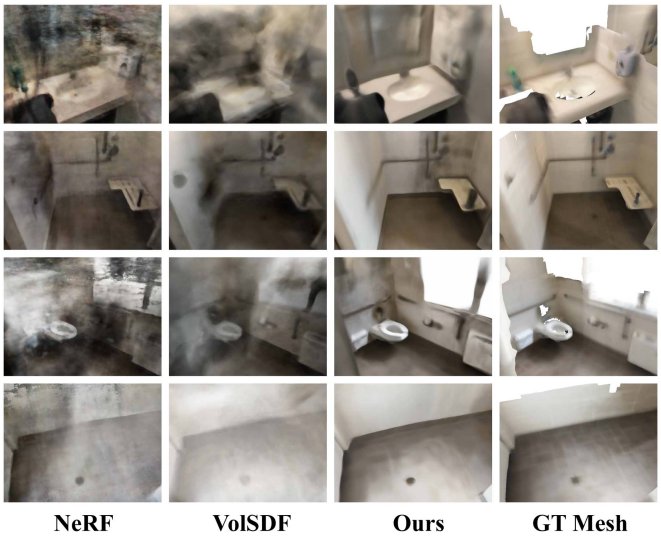


Fig. 10. Novel view synthesis results. We select novel views relatively far from training views for the qualitative comparison. Our method produces better rendering results compared to NeRF and VolSDF. Due to the lack of ground truth images in novel views, we render GT mesh in these views for reference.

missing areas and achieve 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 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 VI. Qualitative results are shown in Fig. 9. 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 is 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 Fig. 9 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 Fig. 10.

VI. Conclusion

In this paper, we introduced a novel indoor scene reconstruction method based on the Manhattan-world and Atlanta-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.

References

[1] M. Atzmon and Y. Lipman, "SAL: Sign agnostic learning of shapes from raw data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2562–2571.
[2] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
[3] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan, "Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 5835–5844.
[4] A. Boulch, B. Le Saux, and N. Audebert, "Unstructured point cloud semantic labeling using deep segmentation networks," in *Proc. Workshop 3D Object Retrieval*, 2017, pp. 17–24.
[5] A. Chen, Z. Xu, A. Geiger, J. Yu, and H. Su, "TensoRF: Tensorial radiance fields," 2022, *arXiv:2203.09517*.
[6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[7] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 833–851.
[8] B. Cheng et al., "SPGNet: Semantic prediction guidance for scene parsing," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 5217–5227.
[9] S. Cheng et al., "Deep stereo using adaptive thin volume representation with uncertainty awareness," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2521–2531.
[10] J. M. Coughlan and A. L. Yuille, "Manhattan world: Compass direction from a single image by Bayesian inference," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, 1999, pp. 941–947.

[11] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "ScanNet: Richly-annotated 3D reconstructions of indoor scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2432–2443.

[12] A. Eftekhar, A. Sax, J. Malik, and A. Zamir, "Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3D scans," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10786–10796.

[13] S. Fridovich-Keil, G. Meanti, F. R. Warburg, B. Recht, and A. Kanazawa, "K-planes: Explicit radiance fields in space, time, and appearance," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 12479–12488.

[14] S. Fridovich-Keil, A. Yu, M. Tancik, Q. Chen, B. Recht, and A. Kanazawa, "Plenoxels: Radiance fields without neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5491–5500.

[15] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski, "Manhattan-world stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1422–1429.

[16] D. Gallup, J.-M. Frahm, and M. Pollefeys, "Piecewise planar and non-planar stereo for urban scene reconstruction," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 1418–1425.

[17] S. J. Garbin, M. Kowalski, M. Johnson, J. Shotton, and J. Valentin, "FastNeRF: High-fidelity neural rendering at 200FPS," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 14346–14355.

[18] A. Gropp, L. Yariv, N. Haim, M. Atzmon, and Y. Lipman, "Implicit geometric regularization for learning shapes," in *Proc. Mach. Learn. Syst.*, 2020, pp. 3569–3579.

[19] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, and P. Tan, "Cascade cost volume for high-resolution multi-view stereo and stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2492–2501.

[20] H. Guo et al., "Neural 3D scene reconstruction with the manhattan-world assumption," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5511–5520.

[21] P. Hedman, P. P. Srinivasan, B. Mildenhall, J. T. Barron, and P. Debevec, "Baking neural radiance fields for real-time view synthesis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 5855–5864.

[22] W. Hu, H. Zhao, L. Jiang, J. Jia, and T.-T. Wong, "Bidirectional projection network for cross dimension scene understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14368–14377.

[23] S. Im, H.-G. Jeon, S. Lin, and I. S. Kweon, "DPSNet: End-to-end deep plane sweep stereo," in *Proc. Int. Conf. Learn. Representations*, 2019.

[24] M. Kazhdan and H. Hoppe, "Screened poisson surface reconstruction," *ACM Trans. Graph.*, vol. 32, 2013, Art. no. 29.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.

[26] A. Kundu et al., "Virtual multi-view fusion for 3D semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 518–535.

[27] U. Kusupati, S. Cheng, R. Chen, and H. Su, "Normal assisted stereo depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2186–2196.

[28] F. J. Lawin, M. Danelljan, P. Tosteberg, G. Bhat, F. S. Khan, and M. Felsberg, "Deep projective 3D semantic segmentation," in *Proc. Int. Conf. Comput. Anal. Images Patterns*, 2017, pp. 95–107.

[29] C. Liu, K. Kim, J. Gu, Y. Furukawa, and J. Kautz, "PlaneRCNN: 3D plane detection and reconstruction from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4445–4454.

[30] L. Liu, J. Gu, K. Z. Lin, T.-S. Chua, and C. Theobalt, "Neural sparse voxel fields," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, Art. no. 1313.

[31] S. Liu, Y. Zhang, S. Peng, B. Shi, M. Pollefeys, and Z. Cui, "DIST: Rendering deep implicit signed distance function with differentiable sphere tracing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2016–2025.

[32] Z. Liu, X. Qi, and C.-W. Fu, "3D-to-2D distillation for indoor scene parsing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4462–4472.

[33] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[34] X. Long, L. Liu, C. Theobalt, and W. Wang, "Occlusion-aware depth estimation with adaptive normal constraints," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 640–657.

[35] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3D surface construction algorithm," in *Proc. 14th Annu. Conf. Comput. Graph. Interactive Techn.*, 1987, pp. 163–169.

[36] P. Merrell et al., "Real-time visibility-based fusion of depth maps," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[37] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3D reconstruction in function space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4455–4465.

[38] B. Mildenhall et al., "NeRF: Representing scenes as neural radiance fields for view synthesis," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 405–421.

[39] T. Müller, "Tiny CUDA neural network framework," 2021. [Online]. Available: https://github.com/nvlabs/tiny-cuda-nn

[40] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," 2022, *arXiv:2201.05989*.

[41] Z. Murez, T. van As, J. Bartolozzi, A. Sinha, V. Badrinarayanan, and A. Rabinovich, "Atlas: End-to-end 3D scene reconstruction from posed images," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 414–431.

[42] R. A. Newcombe et al., "KinectFusion: Real-time dense surface mapping and tracking," in *Proc. 10th IEEE Int. Symp. Mixed Augmented Reality*, 2011, pp. 127–136.

[43] M. Niemeyer, L. Mescheder, M. Oechsle, and A. Geiger, "Differentiable volumetric rendering: Learning implicit 3D representations without 3D supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3501–3512.

[44] M. Oechsle, S. Peng, and A. Geiger, "UNISURF: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 5569–5579.

[45] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "DeepSDF: Learning continuous signed distance functions for shape representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 165–174.

[46] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2019, Art. no. 721.

[47] N. Popovic, D. P. Paudel, and L. Van Gool, "Surface normal clustering for implicit representation of manhattan scenes," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 17860–17870.

[48] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 77–85.

[49] C. R. Qi, H. Su, M. Nießner, A. Dai, M. Yan, and L. J. Guibas, "Volumetric and multi-view CNNs for object classification on 3D data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5648–5656.

[50] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet: Deep hierarchical feature learning on point sets in a metric space," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5105–5114.

[51] C. Reiser, S. Peng, Y. Liao, and A. Geiger, "KiloNeRF: Speeding up neural radiance fields with thousands of tiny MLPs," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 14315–14325.

[52] G. Riegler and V. Koltun, "Free view synthesis," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 623–640.

[53] G. Riegler and V. Koltun, "Stable view synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12211–12220.

[54] A. Romanoni and M. Matteucci, "TAPA-MVS: Textureless-aware patchmatch multi-view stereo," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 10412–10421.

[55] G. Schindler and F. Dellaert, "Atlanta world: An expectation maximization framework for simultaneous low-level edge grouping and camera calibration in complex man-made environments," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2004, pp. I–I.

[56] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4104–4113.

[57] J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys, "Pixelwise view selection for unstructured multi-view stereo," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 501–518.

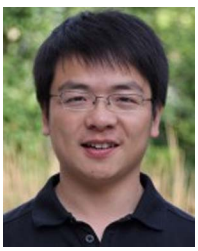
[58] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon, "Scene coordinate regression forests for camera relocalization in RGB-D images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2930–2937.

[59] V. Sitzmann, M. Zollhöfer, and G. Wetzstein, "Scene representation networks: Continuous 3D-structure-aware neural scene representations," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2019, Art. no. 101.

[60] C. Sun, M. Sun, and H.-T. Chen, "Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5449–5459.

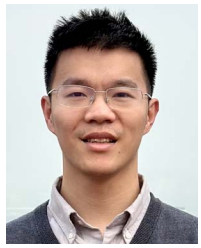
[61] J. Sun, Y. Xie, L. Chen, X. Zhou, and H. Bao, "NeuralRecon: Real-time coherent 3D reconstruction from monocular video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15593–15602.

[62] S. Sun, Y. Zheng, X. Shi, Z. Xu, and Y. Liu, "PHI-MVS: Plane hypothesis inference multi-view stereo for large-scale scene reconstruction," 2021, *arXiv:2104.06165*.
[63] J. Tang, X. Chen, J. Wang, and G. Zeng, "Compressible-composable NeRF via rank-residual decomposition," 2022, *arXiv:2205.14870*.
[64] K. Wadhwani and T. Kojima, "SqueezeNeRF: Further factorized FastNeRF for memory-efficient inference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2716–2724.
[65] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang, "NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 27171–27183 .
[66] Y. Wei, S. Liu, Y. Rao, W. Zhao, J. Lu, and J. Zhou, "NerfingMVS: Guided optimization of neural radiance fields for indoor multi-view stereo," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 5590–5599.
[67] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," 2019. [Online]. Available: https://github.com/facebookresearch/detectron2
[68] Q. Xu and W. Tao, "Planar prior assisted patchmatch multi-view stereo," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12516–12523.
[69] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "MVSNet: Depth inference for unstructured multi-view stereo," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 785–801.
[70] Y. Yao, Z. Luo, S. Li, T. Shen, T. Fang, and L. Quan, "Recurrent MVSNet for high-resolution multi-view stereo depth inference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5520–5529.
[71] L. Yariv, J. Gu, Y. Kasten, and Y. Lipman, "Volume rendering of neural implicit surfaces," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 4805–4815.
[72] L. Yariv et al., "Multiview neural surface reconstruction by disentangling geometry and appearance," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, Art. no. 210.
[73] W. Yin, Y. Liu, C. Shen, and Y. Yan, "Enforcing geometric constraints of virtual normal for depth prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 5683–5692.
[74] A. Yu, R. Li, M. Tancik, H. Li, R. Ng, and A. Kanazawa, "PlenOctrees for real-time rendering of neural radiance fields," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 5732–5741.
[75] Z. Yu, S. Peng, M. Niemeyer, T. Sattler, and A. Geiger, "MonoSDF: Exploring monocular geometric cues for neural implicit surface reconstruction," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 25018–25032.
[76] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6230–6239.
[77] E. Zheng, E. Dunn, V. Jojic, and J.-M. Frahm, "Patchmatch based joint view selection and depthmap estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1510–1517.
[78] S. Zhi, T. Laidlow, S. Leutenegger, and A. J. Davison, "In-place scene labelling and understanding with implicit scene representation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 15818–15827.



Xiaowei Zhou received the PhD degree from The Hong Kong University of Science and Technology, after which he was a postdoctoral researcher with the GRASP Lab, University of Pennsylvania. He is a tenured associate professor of computer science with Zhejiang University, China. His research interests include 3D reconstruction and scene understanding.

Haoyu Guo received the bachelor's degree in computer science from the University of Science and Technology of China, in 2020. He is currently working toward the PhD degree in computer science with Zhejiang University, supervised by Prof. Xiaowei Zhou. His research focuses on 3D scene reconstruction and understanding.



Sida Peng received the PhD degree from the College of Computer Science and Technology, Zhejiang University, in 2023. He is a research professor with the School of Software, Zhejiang University, China. His research interests include 3D reconstruction, rendering, and 3D generation.

Yuxi Xiao is currently working toward the PhD degree with the State Key Laboratory of CAD&CG, Zhejiang University, supervised by Prof. Xiaowei Zhou. His interests lies in 3D vision, graphics and robotics.

Haotong Lin received the bachelor's degree in computer science from Zhejiang University, in 2021. He is currently working toward the PhD degree in computer science with Zhejiang University, advised by Prof. Xiaowei Zhou. His research focuses on 3D/4D reconstruction and object pose estimation.

Qianqian Wang received the bachelor's degree from Zhejiang University, where she worked with Prof. Xiaowei Zhou, and the PhD degree in computer science from Cornell Tech, Cornell University with her advisors Prof. Noah Snavely and Prof. Bharath Hariharan. She is a postdoctoral researcher with UC Berkeley, working with Prof. Angjoo Kanazawa and Prof. Alyosha Efros. Additionally, she is a visiting researcher with Google Research.

Guofeng Zhang received the BS and PhD degrees in computer science from Zhejiang University, in 2003 and 2009, respectively. He is a professor with the State Key Lab of CAD&CG, Zhejiang University. His research interests include structure-from-motion, SLAM, 3D reconstruction, augmented reality, video segmentation and editing.

Hujun Bao received the BSc degree in mathematics from Zhejiang University, in 1987, and the PhD degree in applied mathematics from Zhejiang University, in 1993. He is currently a professor with the State Key Laboratory of CAD&CG and College of Computer Science and Technology, Zhejiang University.