Surface Normal Clustering for Implicit Representation of Manhattan Scenes

Nikola Popovic[1], Danda Pani Paudel[1,2], Luc Van Gool[1,2]

[1]Computer Vision Laboratory, ETH Zurich, Switzerland
[2]INSAIT, Sofia University, Bulgaria

{nipopovic, paudel, vangool}@vision.ee.ethz.ch

Abstract

Novel view synthesis and 3D modeling using implicit neural field representation are shown to be very effective for calibrated multi-view cameras. Such representations are known to benefit from additional geometric and semantic supervision. Most existing methods that exploit additional supervision require dense pixel-wise labels or localized scene priors. These methods cannot benefit from high-level vague scene priors provided in terms of scenes' descriptions. In this work, we aim to leverage the geometric prior of Manhattan scenes to improve the implicit neural radiance field representations. More precisely, we assume that only the knowledge of the indoor scene (under investigation) being Manhattan is known – with no additional information whatsoever – with an unknown Manhattan coordinate frame. Such high-level prior is used to self-supervise the surface normals derived explicitly in the implicit neural fields. Our modeling allows us to cluster the derived normals and exploit their orthogonality constraints for self-supervision. Our exhaustive experiments on datasets of diverse indoor scenes demonstrate the significant benefit of the proposed method over the established baselines. The source code will be available at `https://github.com/nikola3794/normal-clustering-nerf`.

1. Introduction

Above 80% images ever taken are estimated to involve human-made architectural structures, with a substantial share of indoor scenes [11]. These scenes often exhibit strong structural regularities, including flat and texture-poor surfaces in some axis-aligned Cartesian coordinate system – also known as Manhattan world [7, 12]. Paradoxically, these regularities may hinder the 3D modeling process if it is unaware of the human-made scene priors. In fact, several computer vision works have even benefited from the knowledge of the Manhattan world for the task of 3D scene reconstruction [43, 12], camera localization [19], self-calibration [50], and more [37, 35, 29, 57].

For calibrated multi-view cameras, 3D inversion using implicit neural representations [24, 47, 54] is becoming increasingly popular due to their remarkable performance and recent efficiency developments [55, 30, 25, 51]. Meanwhile, such representations are known to benefit from additional supervision in the form of depth [1, 9, 32], normals [18], semantics [45, 58, 17, 15], local-regularization [46, 44, 26], local planar patches [21], or their combinations [56, 14, 21]. In this context, a notable recent work ManhattanSDF [14] demonstrates the benefit of exploiting the high-level geometric prior for structured scenes. More precisely, ManhattanSDF [14] uses the known semantic regions to impose the planar geometry prior of floors and walls under the Manhattan scene assumption. During this process, the exact normal of the floor and partial normals of the walls are assumed to be known, with respect to the camera coordinates.

We aim to improve the 3D neural radiance field representations for calibrated multi-view cameras in indoor Manhattan scenes, with no further assumptions. In other words, we consider that the structural and semantic information is not available, unlike ManhattanSDF [14]. In addition to the floor and walls used in ManhattanSDF [14], we wish to exploit the Manhattan prior of many other indoor scene parts (*e.g.* tables and wardrobes). More importantly, we consider that the Manhattan coordinate frame is also unknown. Our assumptions (of unknown semantics and Manhattan frame) on one hand make our setting very practical. On the other hand, those assumptions make the problem of exploiting the Manhattan scene prior for 3D inversion very challenging.

The virtue of the Manhattan world assumption comes from its simplicity, allowing us to intuitively reason about the geometry of a wide range of complex scenes/objects such as cities, buildings, and furniture. However, such reasoning often requires the axis-aligned Cartesian coordinate frame, also known as the Manhattan frame (MF), to be known [37]. Unfortunately, recovering the Manhattan frame directly from images is not an easy task [7, 37]. Therefore, several methods have been developed until recently [10, 4, 36, 16, 13] to recover the Manhattan frame, relying upon the known 3D reconstruction or image primi-
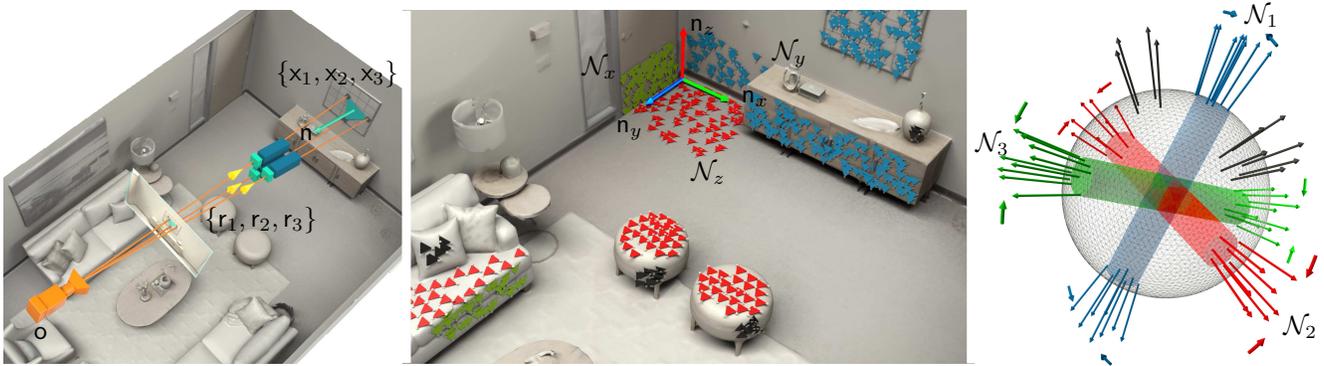
Figure 1: **Illustration of the proposed method.** We compute one surface normal for each ray triplet, using 3D surface points derived from rendered depths (left). The computed normals (from many ray triplets) are clustered to obtain the MF (middle). The non-Manhattan and noisy surfaces are handled by a robust orthogonal normals search, which is later used for self-supervision through the Manhattan prior (right).

tives (eg. lines, planes). We wish to exploit the Manhattan prior for improving the 3D representation, without needing to know MF beforehand. Instead, our experiments reveal that knowing the MF beforehand offers no additional benefit in Manhattan-prior aware radiance field representation.

In this work, we propose a method that jointly learns the Manhattan frame and neural radiance field, from calibrated multi-view in indoor Manhattan scenes, in an end-to-end manner using the recent efficient backbone of Instant-NGP [25]. The proposed method requires no additional information to exploit the Manhattan prior and relies on the explicitly derived normals in the implicit neural fields. We use batches of three neighboring rays, whose effective surface's local piece-wise planarity is assumed to derive the explicit normals by algebraic means. In pure, complete, and enclosed Manhattan scenes, these normals form six clusters corresponding to three orthogonal and other three parallel counterpart planes. However real scenes consist of non-Manhattan scene parts and missing planes. Therefore, we use a robust method that uses minimal three orthogonal clusters to recover the sought Manhattan frame. As in the literature, we seek a rotation matrix whose entries are directly derived from the orthogonal clusters of normals, to align the Manhattan frame. The recovered MF is then used to encourage the derived normals to be axis-aligned for self-supervision. A graphical illustration of our method is presented in Figure 1. Our extensive experiments demonstrate the robustness and utility of our method in improving the implicit 3D in neural radiance field representations.

The major contributions of our paper are listed below:
– We address the problem of exploiting the structured-scene knowledge without requiring any dense or localized scene priors, for the first time in this paper.
– We present a method that successfully exploits the Manhattan scene prior with an unknown Manhattan frame. The proposed method also recovers the unknown frame.
– We demonstrate the robustness and utility of our method

on three indoor datasets, where our method improves the established baselines significantly. These datasets consist of 200+ scenes, making our method tested in significantly more scenes than the state-of-the-art methods.

2. Related Works

Implicit neural representation of 3D: Since the foundational work of *Mildenhall et al.* [24], the implicit neural representation of 3D scenes has advanced on various fronts. These fronts include representation [54, 40], generalization [49, 28, 20, 22, 34], generation [27, 6], and efficient methods [30, 55, 25]. We rely on a recent method Instant-NGP [25] developed by *Mueller et al.*, as our backbone. Our choice is made primarily based on the computational efficiency during both training and inference. Thanks to the offered computational efficiency, we are able to conduct large-scale experiments on several scenes.

Auxiliary supervision methods: In addition to the images, other inputs such as depth [33, 39, 9], semantics [58, 15, 17, 41], normal [52, 53], and their combinations [56, 14, 21] are shown to be beneficial on improving the neural radiance field representation. In this regard, these auxiliary supervisions often use ground-truth labels. It is needless to say that the need for ground-truth supervision is not desired, whenever possible. Therefore, recent methods use labels predicted by some pre-trained networks [56] or recovered from the structure-from-motion (SfM) pipeline [9, 21]. One notable work ManhattanSDF [14] exploits the Manhattan prior without requiring any SfM reconstruction. However, ManhattanSDF requires (a) semantics of the scene parts and (b) the Manhattan frame, to be known. We argue that such labels required for auxiliary supervision cannot always be relied upon, due to domain gaps, poor reconstruction of texture-less regions, and additional computational needs, to list a few. Therefore, we do not use any additional labels for auxiliary supervision.

Manhattan frame estimation: Since the early works of *Bernard* [2], Manhattan structure reasoning is done directly on/from images by detecting the so-called the vanishing-points (VPs) [23, 42]. In fact, the problem of detecting three orthogonal VPs is equivalent to finding the MF in 3D for the calibrated multi-view setting [3]. Note that the knowledge of Manhattan structure has been used in several computer vision works [43, 12, 19, 50, 37, 35, 29, 57]. When unknown, most methods implicitly or explicitly estimate the MF to leverage the Manhattan scene prior. In [37], *Straub et al.* have demonstrated that the MF can be efficiently represented in and recovered from the space of surface normals. We use a similar formulation as [37], using the surface normals derived explicitly from the implicit neural fields.

3. Background and Notations

Manhattan frame (MF) is a coordinate system that is defined by the structure building orthogonal planes of Manhattan scenes. We consider unknown MF since the scene planes and their geometric relationships are unknown. Let the unknown MF differs from the world-frame (WF), used for the 3D representation, by rotation $R \in SO(3)$. We denote three orthogonal axes in MF by $\mathcal{E} = \{e_x, e_y, e_z\}$. Without loss of generality, let the axes' coordinates be $e_x = [1, 0, 0]^\mathsf{T}$, $e_y = [0, 1, 0]^\mathsf{T}$, and $e_z = [0, 0, 1]^\mathsf{T}$. Note that these axes align with the normals of the respective scene building planes, in the WF. Let $\mathcal{N} = \{n_i\}_{i=1}^m$ be a set of 3D normals of all the scene planes. Then, the rotation R, from world to Manhattan frame, aligns the normals $n \in \mathcal{N}$ to the Manhattan axes $e \in \mathcal{E}$, i.e, $Rn_i \in \mathcal{E}, \forall n_i \in \mathcal{N}$.

We are interested to recover R from a set of normals $\mathcal{N}$. For this, the above set-to-set assignment alone is not sufficient. This requires the element-wise assignment between sets $\mathcal{N}$ and $\mathcal{E}$. To do so, we divide the set $\mathcal{N}$ into three orthogonal subsets $\mathcal{N}_x$, $\mathcal{N}_y$, and $\mathcal{N}_z$. Now, for any triplet of $\{n_x, n_y, n_z\}$ from the corresponding orthogonal subsets, we aim to establish the condition $[e_x, e_y, e_z] = R[n_x, n_y, n_z]$. Note that the assignment condition results into $R = [n_x, n_y, n_z]^\mathsf{T}$. Therefore, the problem of recovering MF from a given set of normals boils down to finding three normals from orthogonal subsets. At this point, one issue regarding robustness remains pending. More precisely, we wish to recover R for a noisy set of normals $\mathcal{N}$, with potentially overwhelmingly many outliers.

We use a robust method to recover R from given normals $\mathcal{N}$. In principle, R can be recovered from minimal two orthogonal normals, with one additional normal for disambiguation by sign correction. The recovered rotation can then be validated by consensus for robust recovery [37]. Alternatively, one can also perform the robust estimation of the orthogonal subsets of $\mathcal{N}$, followed by solving $[e_x, e_y, e_z] = R[n_x, n_y, n_z], \forall n_x \in \mathcal{N}_x, n_y \in \mathcal{N}_y, n_z \in \mathcal{N}_z$, for $R \in SO(3)$. For computational reasons, we estimate

the robust subsets by clustering. The orthogonal subsets are then obtained by choosing the three most orthogonal clusters. Later, the obtained cluster centers are used to estimate the MF (or to enforce its existence) represented by R. Please refer to Section 4.2 for more details.

4. Method

From a set of calibrated images, we model the 3D scene using the neural radiance field representation. In the process, we wish to exploit the Manhattan scene prior, without the knowledge of the Manhattan frame. This problem is addressed by jointly optimizing for the neural radiance field and the Manhattan frame estimation. The complete pipeline of our method is illustrated in Figure 2. As shown, our method consists of three units: (a) Explicit normal modeling, (b) Robust estimation of orthogonal normals, and (3) Self-supervision by Manhattan prior. In the first stage, we shoot a batch of three rays which allows us to estimate the surface normal using an algebraic method. As the estimated normals are bound to be noisy (with outliers), we perform their clustering to obtain the most orthogonal clusters (representing the Manhattan frame) in a robust manner. The obtained clusters are then used to estimate the sought MF, in the form of a rotation matrix, using the method discussed in Section 3. In the final step, we use the estimated Manhattan frame to encourage the close-by normals to be Manhattan-like and to enforce a stricter orthogonality constraint for self-supervision. In the following, we will present the details of three units of our method in different Subsections.

4.1. Explicit Normal Modelling

At any given view, we consider the color $c = C_\theta(r)$ and depth $d = D_\theta(r)$ for a ray r emanating from the corresponding camera center o is obtained using the volumetric rendering of the implicit neural radiance field, represented by a neural network parameterized by θ as in [24]. Then, the location of the 3D surface point is given by, $x = o + d.r$. We process triplets of three neighboring rays. Let $\mathcal{T} = \{r_1, r_2, r_3\}$ be such a triplet, whose corresponding surface points are given by $\mathcal{X} = \{x_1, x_2, x_3\}$. Now, for the ray triplet $\mathcal{T}$, we obtain the explicit surface normal n with the help of the point triplet $\mathcal{X}$, using the following mapping and algebraic operation,

$$\mathcal{T} \rightarrow \mathcal{X} \rightarrow v = (x_1 - x_2) \times (x_1 - x_3) \rightarrow n = \frac{\text{sign}(o^\mathsf{T}v)v}{\|v\|}. \quad (1)$$

Note that v is a vector orthogonal to the plane passing through 3D points in triplet $\mathcal{T}$. We obtain the oriented normal n by normalizing v and correcting its sign by ensuring that the camera center o is in front of the estimated plane. A graphical illustration of the surface normal estimation is provided in Figure 1 on the left. We select a random set
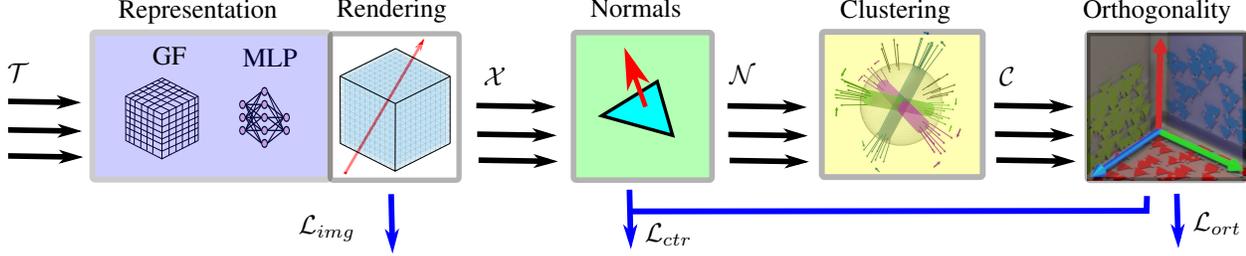
Figure 2: **The complete pipeline of our method.** We use grid features (GF) and an MLP to represent the radiance field. The explicit normals are derived using the depths obtained from volume rendering. The Manhattan scene prior is exploited by clustering the estimated normals to enforce their orthogonality.

of ray triplets $\{\mathcal{T}_i\}_{i=1}^m$ from multiple cameras. These ray triplets provide us the surface point triplets $\{\mathcal{X}_i\}_{i=1}^m$. We use point triplets to explicitly derive the surface normals $\mathcal{N} = \{n_i\}_{i=1}^m$, using (1). These normals are later used to recover the unknown Manhattan frame. The simplicity of the explicit normals computed in this paper makes them easy to compute and handle. If needed, normals of different sizes of surface regions could also be computed similarly.

4.2. Robust Estimation of Orthogonal Normals

We are interested to recover the MF from a set of noisy surface normals $\mathcal{N} = \{n_i\}_{i=1}^m$. Recall Section 3, the MF can be obtained by robustly recovering a set of three orthogonal normals from $\mathcal{N}$. To do so, we first cluster $\mathcal{N}$ into k clusters $\{\mathcal{C}_i\}_{i=1}^k$, by using the the well-known k-means clustering algorithm [5]. During the clustering, every centroid is L_2 normalized after each iteration, to ensure that they are unit vectors to represent surface normals. In a perfect Manhattan world, there exist only six clusters corresponding to the three orthogonal MF axes $\mathcal{E} = \{e_x, e_y, e_z\}$ and their parallel counterparts. However, real scenes consist of surfaces of different orientations. An additional source of non-Manhattan normals comes from the inaccuracy in the normal estimation. Therefore, we use a clustering technique to robustly recover the orthogonal normals. In the following, we present how the orthogonal clusters are selected from the set of clusters $\{\mathcal{C}_i\}_{i=1}^k$. Note from Section 3, the selected orthogonal clusters are considered to represent MF defining sets $\mathcal{N}_x$, $\mathcal{N}_y$, and $\mathcal{N}_z$. We proceed by first selecting three orthogonal clusters, say $\mathcal{N}_1$, $\mathcal{N}_2$, and $\mathcal{N}_3$. Later, we assign them to $\mathcal{N}_x$, $\mathcal{N}_y$, and $\mathcal{N}_z$ to recover MF.

For notational ease, we pair clusters and centroids as $\mathcal{U} = \{(\mathcal{C}_i, c_i)\}_{i=1}^k$, where the centroids are computed by taking the average across the corresponding cluster such that $c_i = \frac{1}{|\mathcal{C}_i|} \sum_{n \in \mathcal{C}_i} n$, followed by normalizing to a unit vector $c_i = \frac{c_i}{\|c_i\|}$. Leveraging the Manhattan prior through the assumption that Manhattan surfaces dominate the scene, we pick $n_1 = c_f$, where $f = \arg\max_i |\mathcal{C}_i|$. In other words, we pick the centroid of the largest cluster, as one of the MF axes. Then, we obtain $n_2 = c_s$ and $n_3 = c_t$ as a solution of the optimization problem

---

Algorithm 1 $(\mathcal{N}_1, \mathcal{N}_2, \mathcal{N}_3, R) = \text{findManhattanFrame}(\mathcal{N})$

1. Cluster normals $\mathcal{N}$ into, $\mathcal{U} = \{(\mathcal{C}_i, c_i)\}_{i=1}^k$ using k-means.
2. For the largest cluster $\mathcal{C} \in \mathcal{U}$, assign $(\mathcal{C}, c) \to (\mathcal{N}_1, n_1)$.
3. Find $\mathcal{C}_s \in \mathcal{U}, \mathcal{C}_t \in \mathcal{U}$ minimizing $|c_s^\mathsf{T} n_1| + |n_1^\mathsf{T} c_t| + |c_s^\mathsf{T} c_t|$.
4. Assign $(\mathcal{C}_s, c_s) \to (\mathcal{N}_2, n_2), (\mathcal{C}_t, c_t) \to (\mathcal{N}_3, n_3)$.
5. $|e_z^\mathsf{T} n_1| \leq \frac{1}{\sqrt{2}}? n_1 \to n_z : (|e_y^\mathsf{T} n_1| \leq \frac{1}{\sqrt{2}}? n_1 \to n_y : n_1 \to n_x)$.
6. Perform the remaining $\{n_1, n_2, n_3\} \to \{n_x, n_y, n_z\}$ as in step 5 to the closest corresponding canonical axes.
7. Return $\mathcal{N}_1, \mathcal{N}_2, \mathcal{N}_3$, and $R = [n_x, n_y, n_z]^\mathsf{T}$.

$s, t = \arg\min_{i,j} |c_i^\mathsf{T} n_1| + |n_1^\mathsf{T} c_j| + |c_i^\mathsf{T} c_j|$. In other words, we find two additional cluster centroids providing the most orthogonal triplet. We further merge all selected clusters with their opposites. This is achieved by comparing all cluster pairs. Whenever two centroids are nearby, but opposite in sign, the corresponding clusters are merged with the appropriate sign correction. The procedure of finding $\{\mathcal{N}_1, \mathcal{N}_2, \mathcal{N}_3\}$ is illustrated in Figure 1 (middle and right), and is summarized in Algorithm 1 from step 1–4. Although the orthogonal clusters with their centroids are sufficient to exploit the desired Manhattan prior, we may wish to recover the MF in the form of a rotation matrix. As such, any arrangement of $\{n_1, n_2, n_3\}$ as a valid rotation matrix offers us a valid MF. We may however often be interested to recover the one which is closest to the world frame. For this reason, we suggest keeping the largest cluster's centroid paired to the closes canonical axis in $\mathcal{E}$. Similarly, we also align one more axis, whereas the last remaining axis gets paired by default. We summarize how to recover MF closest to the world frame in Algorithm 1 in steps 5–7.

4.3. Self-supervision by Manhattan Prior

The supervision of the implicit neural radiance field, parameterized by θ, is achieved through jointly optimizing three loss terms. The first one is the photometric loss computed as follows, $\mathcal{L}_{img} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \|C_\theta(r) - C(r)\|_2^2$, where $C(r)$ is the ground-truth color, and $\mathcal{R}$ is the set of rays going through sampled pixel triplets. This loss term is responsible for facilitating the learning of the implicit 3D representation of the investigated scene [24].

Losses from Manhattan prior: We exploit the Manhattan

scene prior for self-supervision, in order to improve the implicit 3D representation without additional ground truth labels. We do so by using the clusters $\{\mathcal{N}_1, \mathcal{N}_2, \mathcal{N}_3\}$ obtained from Algorithm 1. More precisely, for cluster-centroid pairs $\{(\mathcal{N}_i, \mathsf{n}_i)\}_{i=1}^3$ we use the following two losses,

$$\mathcal{L}_{ctr} = \frac{1}{3}\sum_{i=1}^3 \frac{1}{|\mathcal{N}_i|}\sum_{\mathsf{n}\in\mathcal{N}_i}\|1 - \mathsf{n}_i^\mathsf{T}\mathsf{n}\|_1 + \|\mathsf{n}_i - \mathsf{n}\|_1, \quad (2)$$

$$\mathcal{L}_{ort} = \frac{1}{3}(|\mathsf{n}_1^\mathsf{T}\mathsf{n}_2| + |\mathsf{n}_1^\mathsf{T}\mathsf{n}_3| + |\mathsf{n}_2^\mathsf{T}\mathsf{n}_3|). \quad (3)$$

Here, the loss $\mathcal{L}_{ctr}$ encourages tighter clusters, while the loss $\mathcal{L}_{ort}$ enforces the orthogonality among the three clusters. The final loss used to optimize θ is then given by $\mathcal{L} = \mathcal{L}_{img} + \lambda_{ctr}\mathcal{L}_{ctr} + \lambda_{ort}\mathcal{L}_{ort}$, where λ_{ctr} and λ_{ort} are the hyperparameters. Please, refer to Figure 2 for a schematic summary of all three losses.

5. Experiments

5.1. Baselines, Metrics, and Implementation Details

InstantNGP [25] (baseline): This method represents the scene as a multi-resolution voxel grid and leverages a hash table of trainable feature vectors, which are used to represent grid elements. The representation is further processed with a small MLP. This allows InstantNGP to be very computationally efficient, requiring in our case around 30 minutes per scene to train and evaluate. Therefore, we use it as the backbone of our method and as our baseline.

ManhattanDF [14]: This method exploits the Manhattan prior by supervising the explicit normals of the floors to align with the known n_z, as well as by supervising normals of the walls to align with two learned orthogonal axes which are also orthogonal to n_z. To achieve this, apart from the RGB images, this method relies on knowing the wall and floor semantics, as well as knowing the exact floor axis n_z (MF partially known). We implemented it on top of the InstantNGP backbone with density field estimation.

Ours: We implement our proposed method from Section 4 on top of InstantNGP with density field estimation.

Ours (MF known): We modify our method by assuming that the MF is fully known. We do so by adding additional loss terms $\mathcal{L}_{man_i} = \|1 - \mathsf{n}_i^\mathsf{T}\mathsf{m}\|_1 + \|\mathsf{n}_i - \mathsf{m}\|_1$ for each orthogonal cluster centroid n_i and its closest MF axis m (or a closer opposite counterpart). Thus, we explicitly guide the orthogonal triplet to align with the known MF.

Metrics: To quantitatively evaluate novel view rendering, we use peak signal-to-noise ratio (PSNR) and the structural similarity index (SSIM) [48]. To evaluate the extracted surface normals of novel rendered views, we use the median angular error. To partially evaluate the quality of the learned implicit 3D structure, we utilize the mean absolute error (MAE) and the root mean square error (RMSE) on the

depth obtained by volume rendering. Finally, to evaluate the recovered MF, we calculate the absolute error between the yaw, pitch, and roll angles of the recovered frame and the MF. All metrics are averaged across scenes, except for the yaw, pitch, and roll errors for which the median is reported.

Implementation Details: We turn on $\mathcal{L}_{ort}$ and $\mathcal{L}_{centr}$ after 500 steps and linearly increase their weights to the specified values over the next 2500 steps. Also, we randomly sample rays for one-third of every batch size and select their left and upper neighbor to form a triplet to facilitate obtaining explicit surface normals. For other implementation details, please refer to the supplementary material.

Note that in addition to RGB images, ManhattanDF [14] requires knowing the wall and floor semantics, and the exact floor axis n_z. In contrast, our method only requires RGB images and the assumed Manhattan prior to holding true. Therefore, this comparison aims to get a better insight into what can be achieved without leveraging additional labels.

5.2. Datasets

Hypersim [31] is a photorealistic synthetic dataset consisting of indoor scenes. It was created by leveraging a large repository of scenes created by professional artists, with 461 indoor scenes in total. This dataset is geometry-rich, containing a lot of details and lighting sources. Each scene has one or more camera trajectories available, where each trajectory has up to 100 views rendered in 1024×768. For each scene, camera calibration information is provided, as well as detailed per-pixel labels such as depth and surface normals. After cleaning up the scenes by discarding a few with insufficient camera views, as well as other problems, we were left with 435 scenes. We also kept only one camera trajectory per scene. We then evaluated the InstantNGP [25] baseline on all 435 scenes and made three divisions based on its PSNR on unseen views. **Hypersim-A** contains 20 scenes where the baseline was performing well above average, **Hypersim-B** contains 20 scenes where the baseline had near-average performance, and **Hypersim-C** contains 10 scenes where the baseline had a below-average performance. For each scene, we randomly assigned half views to the training split and the rest to the test split. We note that the test split often contains views that were partially or completely unobserved during training. We also note that Hypersim is visually very realistic, geometry-rich, and challenging, considering it is synthetic. This can be subjectively observed by inspecting rendered views.

ScanNet [8] is a real-world dataset consisting of indoor scenes. It was collected using a scalable RGB-D capture system that includes automated surface reconstruction and crowd-sourced semantic annotation. The dataset contains 1613 indoor scenes, which are annotated with ground-truth camera poses, surface reconstructions, and instance-level semantic segmentations. We use three scenes that were used

Table 1: **Experiments on Hypersim.** We observe that our method consistently outperforms the baseline, as well as the ManhattanDF. This is very interesting since, unlike ManhattanDF, we do not use any additional labels during training. Finally, we see that it does not matter for our method whether the MF is known beforehand or not. Therefore, the additional knowledge of MF is neither necessary nor helpful.

		PSNR↑	SSIM↑	Normals° ↓	Pitch° ↓	Roll° ↓	Yaw° ↓	D-MAE↓	D-RMSE↓
Scenes A	InstantNGP [25] (baseline)	25.86	0.871	57.12	6.18	6.46	19.25	0.064	0.102
	ManhattanDF [14]	26.51	0.868	40.69	1.23	1.01	5.08	0.053	0.087
	Ours	27.20	0.864	37.30	0.40	0.50	0.52	0.053	0.093
	Ours (MF known)	27.21	0.856	35.59	0.25	0.26	0.45	0.052	0.091
Scenes B	InstantNGP [25] (baseline)	20.75	0.811	60.12	6.06	7.92	15.87	0.105	0.151
	ManhattanDF [14]	21.87	0.826	50.50	2.55	2.06	11.69	0.079	0.121
	Ours	22.45	0.816	54.08	1.19	1.35	1.81	0.080	0.127
	Ours (MF known)	22.51	0.813	50.59	0.51	0.65	0.55	0.078	0.126
Scenes C	InstantNGP [25] (baseline)	17.79	0.740	64.29	7.45	4.55	15.14	0.130	0.174
	ManhattanDF [14]	18.33	0.770	56.08	3.20	3.41	10.25	0.103	0.147
	Ours	19.43	0.768	54.79	5.37	2.24	4.24	0.094	0.133
	Ours (MF known)	19.29	0.764	55.12	3.64	4.03	9.48	0.094	0.135
194 scenes	InstantNGP [25] (baseline)	20.47	0.783	61.34	6.56	6.99	21.75	0.104	0.146
	ManhattanDF [14]	20.94	0.794	52.81	1.72	2.32	13.48	0.097	0.140
	Ours	21.63	0.786	52.01	1.87	1.94	3.77	0.085	0.126

in [14], where one-tenth of all views were uniformly sampled, leaving 303-477 views per scene. For each scene, the training and test split both contain half of the total views.

Replica [38] is a synthetic dataset featuring a diverse set of 18 indoor scenes. Each scene is equipped with photorealistic textures, allowing one to render realistic images from arbitrary camera poses. We use five scenes that were rendered and prepared in [58], where the rendered views also contain semantic segmentation labels. Each scene contains 900 views generated from random 6-DOF trajectories similar to hand-held camera motions in 640x480 resolution. We select 75 views for training and 75 for testing in each scene.

5.3. Results

Experiments on Hypersim: We summarize the experiments performed on Hypersim in Table 1. Our proposed method achieves clear improvements in comparison to the InstantNGP baseline, in terms of novel-view rendering, normals estimation, depth estimation, as well as MF recovery. This is consistent across different scene difficulties, namely splits A, B, and C. Furthermore, our method also outperforms the ManhattanDF SotA. This is very interesting since, unlike ManhattanDF, we do not use any additional labels during training, other than the ground truth RGB. The ManhattanDF has partial access to ground truth MF axes, as well as wall and floor semantics, which directly implies having sparse surface normals ground truth. However, there are lots of Manhattan objects and areas in realistic scenes such as in Hypersim, other than the walls and floors. Our self-supervised method leverages the presence of many such objects and areas, to facilitate imposing geometrical constraints on the implicit representation during learning. This can be visually observed in Figure 3. Moreover, we observe that it does not matter for our method whether the MF is known beforehand, since our method recovers the

(a) InstantNGP [25] (baseline) (b) ManhattanDF [14]

(c) Ours (d) Grond truth
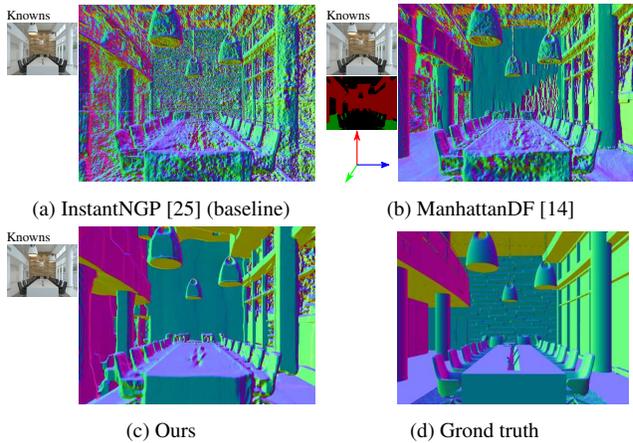
Figure 3: **Leveraging the Manhattan prior (Hypersim-A).** Our method leverages the presence of many Manhattan objects and areas in the scene in a self-supervised fashion, which facilitates imposing geometrical constraints during learning. Our method offers plausible normals – sometimes with missing details – that help to better model the radiance fields. In contrast, ManhattanDF leverages the Manhattan prior through labels of only walls and floors.

MF in the clustering step automatically, and therefore similar performance is achieved in both cases. This is particularly exciting because our self-supervised method performs similarly to the supervised one (using the ground-truth MF). Finally, we show large-scale experiments on a larger set of 194 scenes, which lead to the same conclusions. We observed that the semantic loss of the ManhattanDF sometimes causes convergence issues on difficult scenes, which we partially alleviated with class weighting and label smoothing. Therefore, we report results on only 194 scenes where ManhattanDF converged. Additional results on all 435 scenes are provided in the supplementary.

Experiments on ScanNet: In Table 2, we examine the behavior of our proposed method on real-world indoor scenes.
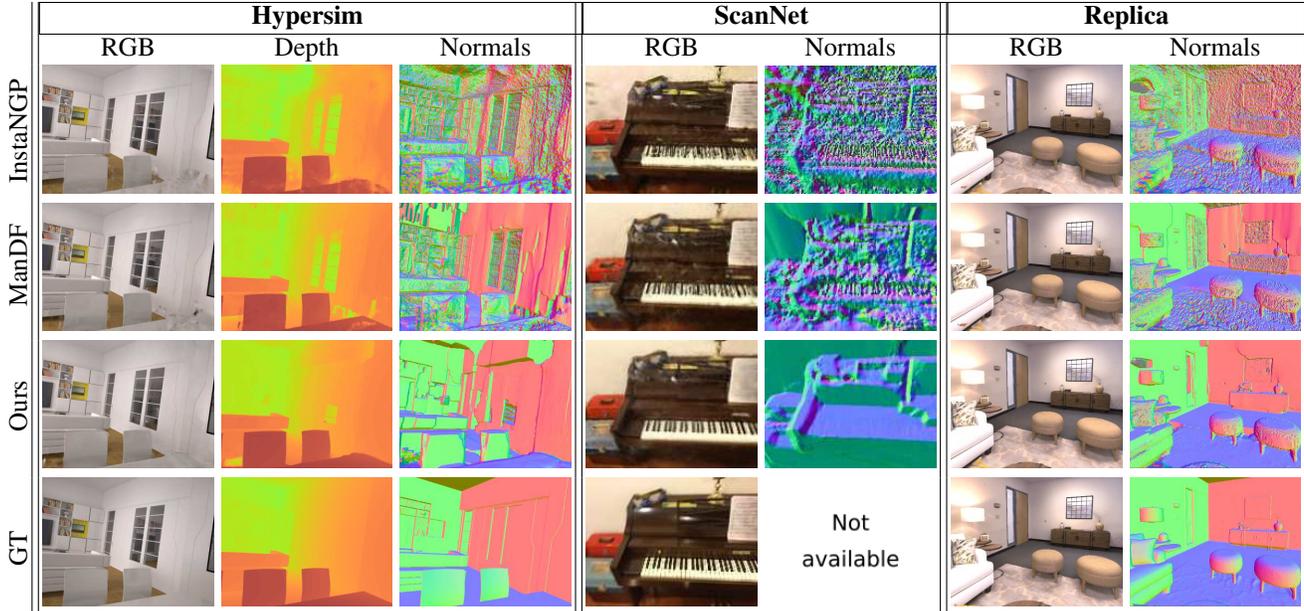
Figure 4: **Qualitative results.** Our method leverages many Manhattan objects and surfaces in the scene, which improves the implicit geometrical representation compared to the baseline. Unlike ManhattanDF, our method relies on many cues other than the walls & floors, which leads to a better representation of some objects (e.g. the white table in the bottom right corner of the example from columns 1-3). For more qualitative results, please refer to the supplementary material.

Table 2: **Experiments on real-world ScanNet data.** Our method outperforms both the baseline and ManhattanDF. This is the case both with and without supervising with sparse depth from SfM.

	PSNR↑	SSIM ↑	Depth↓-MAE
InstantNGP [25] (baseline)	17.78	0.587	0.119
RegNeRF [26]	18.73	0.618	0.102
ManhattanDF [14]	18.68	0.614	0.112
Ours	20.79	0.643	0.072
+ additional sparse depth from SfM			
InstantNGP [25] (baseline)	20.70	0.631	0.048
ManhattanDF [14]	21.53	0.640	0.052
Ours	22.25	0.667	0.033

Table 3: **Experiments on Replica.** Our method outperforms the baseline, and it shows similar performance as ManhattanDF, which leverages additional labels during training.

	PSNR↑	SSIM ↑	Depth↓-MAE
InstantNGP [25] (baseline)	34.30	0.944	0.022
Semantic-NeRF [58]	34.08	0.938	0.014
ManhattanDF [14]	35.24	0.944	0.008
Ours	35.13	0.944	0.011

Our method achieves clear improvements in comparison to the InstantNGP baseline, as well as to ManhattanDF, in terms of all measured metrics. Additionally, following the experimental setting of ManhattanDF [14], we train all methods with additional sparse depth supervision – where the sparse depth is obtained from the SfM pipeline. Our proposed method is also superior in this setting.

Experiments on Replica: We summarize experiments performed on the Replica dataset in Table 3. It can be observed that the InstantNGP baseline already performs very
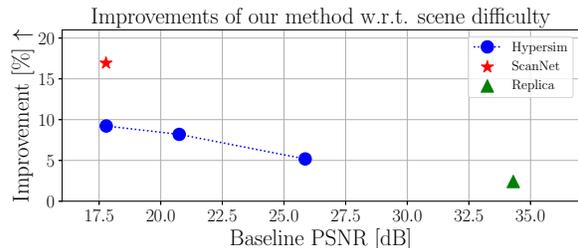
Figure 5: **Improvements vs. scene difficulty.** Our method achieves the biggest improvements on scenes of hard and moderate difficulty, by leveraging many Manhattan objects and surfaces.

well on this dataset, leaving not much room for further improvements. Therefore, we consider Replica as an easier dataset. Nevertheless, our method still performs better than the baseline. As expected, ManhattanDF also improves the baseline similarly. Recall that the ManhattanDF uses additional labels for supervision. We also note that Replica has noticeably more walls and floors, and less of other objects, compared to other more complex datasets.

Qualitative results: We depict the visual results of the discussed experiments in Figure 4. Our method leverages many Manhattan objects and surfaces in the scene, which improves the geometrical structure of 3D compared to the InstantNGP baseline. This is visible in surface normals and depth, obtained using volume rendering. Furthermore, we observe that, unlike ManhattanDF, our method relies on many Manhattan cues other than the walls & floors. This leads to a better representation of some such objects, e.g. the white table in the bottom right corner of the example

Table 4: **Sparse training views.** Our proposed method clearly exhibits the best performance, for all cases of input view sparsity.

		PSNR↑	SSIM↑	Depth↓-MAE
12 views	InstantNGP [25]	18.02	0.706	0.138
	ManhattanDF [14]	19.45	0.750	0.116
	Ours	20.50	0.760	0.104
9 views	InstantNGP [25]	16.79	0.661	0.154
	ManhattanDF [14]	18.04	0.714	0.130
	Ours	19.14	0.728	0.120
6 views	InstantNGP [25]	15.75	0.582	0.178
	ManhattanDF [14]	16.00	0.639	0.159
	Ours	16.67	0.667	0.158

Table 5: **Ablation study on Hypersim-A.** Both of our proposed loss terms contribute to the overall performance.

	PSNR↑	Norm.°↓	Depth↓	Rot.°↓
Only $\mathcal{L}_{img}$	25.86	57.12	0.064	10.63
+ $\mathcal{L}_{ort}$	27.21	50.07	0.058	5.39
+ $\mathcal{L}_{ctr}$	27.06	36.09	0.053	0.57
+ $\mathcal{L}_{ort}$ + $\mathcal{L}_{ctr}$ (Ours)	27.20	37.30	0.053	0.47
Ours + MF known	27.21	35.59	0.052	0.32
Ours (no delay)	27.06	35.78	0.054	0.39
Ours (no w lin.)	27.01	37.86	0.055	0.60

from columns 1-3. Moreover, we observe that our method is able to cope with difficult scenes and views, where other methods struggle. For more qualitative results, please refer to the supplementary material.

Improvements with respect to scene difficulty: We analyze the improvements by our method for different scene difficulties. We decided on scene difficulty based on the novel-view rendering performance of the InstantNGP baseline. In Figure 5, we see that our method brings the most benefits for scenes of hard and moderate difficulties, thanks to the Manhattan scene prior.

Sparse training views: In order to gain more insight, we examine the behavior of training neural radiance fields with sparse training input views. The results on the Hypersim-A dataset are presented in Table 4. Our proposed method clearly outperforms the InstantNGP baseline, as well as ManhattanDF, when trained with 6, 9, and 12 input views.

Finding the MF: We test the robustness of our proposed method for finding the Manhattan frame, by introducing a simultaneous offset α in the yaw, pitch, and roll on the canonical MF. The experiments are performed on Hypersim A and can be found in Figure 6. In Figure 6a we observe that the novel-view rendering quality remains largely robust to the rotation offset α. We note that we increase the scene bounding box by the same factor for all experiments in Figure 6, to make sure that objects remain within the voxel grid for maximal α. This slightly decreases the resolution of grid element representations, so PSNR is slightly lower than in Table 1. Furthermore, in Figure 6b we see that the MF estimation remains robust to the rotation offset.

Ablation study: We report our ablation study in Table 5. Both of our proposed losses contribute to the overall performance. Furthermore, turning on $\mathcal{L}_{ort}$ and $\mathcal{L}_{centr}$ after

(a) Novel-view rendering quality remains robust to the rotation offset α.

(b) MF estimation remains robust to the rotation offset α.
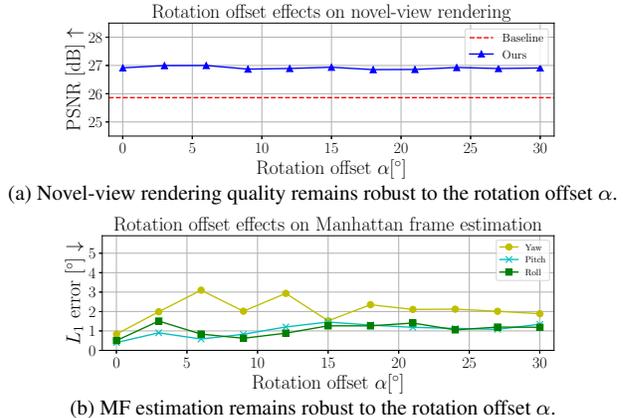
Figure 6: **Finding the MF.** We test the robustness of our method by introducing the rotation offsets on the canonical MF.

500 steps, and linearly increasing their weight also helps slightly. Please refer to the supplementary for more details.

6. Conclusion

We demonstrated the possibility of exploiting the Manhattan scene prior without needing any additional supervision. This is achieved by performing robust clustering of explicit normals, followed by the search of the Manhattan frame (MF) whose existence is based on the assumed prior. The sought MF is obtained from the orthogonal clusters, which are later used to self-supervise the neural representation learning. Our self-supervision encourages the normals of Manhattan surfaces to group into three orthogonal directions. Our experiments on three indoor datasets demonstrate that the proposed method not only benefits from building parts (such as walls and floors) but also exploits many other Manhattan scene parts (such as furniture). Both quantitative and qualitative evaluations reveal the benefit of the proposed method in terms of, improved performance over the established baselines and competitive results to state-of-the-art methods that use additional labels for supervision. Our method has the potential to be extended in other higher expressiveness scene priors, such as Atlanta world and the mixture of Manhattan frames.

Limitations: One limitation of our method is that it sometimes produces surface normals with missing details or "blocky" artifacts. Nevertheless, this usually offers better novel-view RGB rendering, compared to not imposing any structure priors. Another limitation is that our method is not beneficial for very easy scenes. For a detailed discussion of limitations, please refer to the supplementary material.

Acknowledgments

References

[1] Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural rgb-d surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6290–6301, 2022.

[2] Stephen T Barnard. Interpreting perspective images. *Artificial intelligence*, 21(4):435–462, 1983.

[3] Jean-Charles Bazin, Hongdong Li, In So Kweon, Cédric Demonceaux, Pascal Vasseur, and Katsushi Ikeuchi. A branch-and-bound approach to correspondence and grouping problems. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1565–1576, 2012.

[4] Jean-Charles Bazin, Yongduek Seo, Cédric Demonceaux, Pascal Vasseur, Katsushi Ikeuchi, Inso Kweon, and Marc Pollefeys. Globally optimal line clustering and vanishing point estimation in manhattan world. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 638–645. IEEE, 2012.

[5] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[6] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022.

[7] James M Coughlan and Alan L Yuille. Manhattan world: Compass direction from a single image by bayesian inference. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 941–947. IEEE, 1999.

[8] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

[9] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12882–12891, 2022.

[10] Patrick Denis, James H Elder, and Francisco J Estrada. Efficient edge-based methods for estimating manhattan frames in urban imagery. In *European conference on computer vision*, pages 197–210. Springer, 2008.

[11] Clara Fernandez-Labrador. Indoor scene understanding using non-conventional cameras. *PhD thesis*, 2020.

[12] Yasutaka Furukawa, Brian Curless, Steven M Seitz, and Richard Szeliski. Manhattan-world stereo. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1422–1429. IEEE, 2009.

[13] Wuwei Ge, Yu Song, Baichao Zhang, and Zehua Dong. Globally optimal and efficient manhattan frame estimation by delimiting rotation search space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15213–15221, 2021.

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

[15] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5885–5894, 2021.

[16] Kyungdon Joo, Tae-Hyun Oh, Junsik Kim, and In So Kweon. Globally optimal manhattan frame estimation in real-time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1763–1771, 2016.

[17] Abhijit Kundu, Kyle Genova, Xiaoqi Yin, Alireza Fathi, Caroline Pantofaru, Leonidas J Guibas, Andrea Tagliasacchi, Frank Dellaert, and Thomas Funkhouser. Panoptic neural fields: A semantic object-aware neural scene representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12871–12881, 2022.

[18] Fu Li, Hao Yu, Ivan Shugurov, Benjamin Busam, Shaowu Yang, and Slobodan Ilic. Nerf-pose: A first-reconstruct-then-regress approach for weakly-supervised 6d object pose estimation. *arXiv preprint arXiv:2203.04802*, 2022.

[19] Haoang Li, Jian Yao, Jean-Charles Bazin, Xiaohu Lu, Yazhou Xing, and Kang Liu. A monocular slam system leveraging structural regularity in manhattan world. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2518–2525. IEEE, 2018.

[20] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5741–5751, 2021.

[21] Zhi-Hao Lin, Wei-Chiu Ma, Hao-Yu Hsu, Yu-Chiang Frank Wang, and Shenlong Wang. Neurmips: Neural mixture of planar experts for view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15702–15712, 2022.

[22] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7210–7219, 2021.

[23] Gerard F McLean and D Kotturi. Vanishing point detection by line clustering. *IEEE Transactions on pattern analysis and machine intelligence*, 17(11):1090–1095, 1995.

[24] 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, 2020. cite arxiv:2003.08934Comment: ECCV 2020 (oral). Project page with videos and code: http://tancik.com/nerf.

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

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

[27] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021.

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

[29] Pulak Purkait, Christopher Zach, and Ales Leonardis. Rolling shutter correction in manhattan world. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 882–890, 2017.

[30] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14335–14345, 2021.

[31] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *International Conference on Computer Vision (ICCV) 2021*, 2021.

[32] Barbara Roessle, Jonathan T Barron, Ben Mildenhall, Pratul P Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12892–12901, 2022.

[33] Antoni Rosinol, John J Leonard, and Luca Carlone. Nerf-slam: Real-time dense monocular slam with neural radiance fields. *arXiv preprint arXiv:2210.13641*, 2022.

[34] Mehdi SM Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lučić, Daniel Duckworth, Alexey Dosovitskiy, et al. Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6229–6238, 2022.

[35] Gautam Singh and Jana Kosecka. Visual loop closing using gist descriptors in manhattan world. In *ICRA omnidirectional vision workshop*, pages 4042–4047, 2010.

[36] Julian Straub, Nishchal Bhandari, John J Leonard, and John W Fisher. Real-time manhattan world rotation estimation in 3d. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1913–1920. IEEE, 2015.

[37] Julian Straub, Oren Freifeld, Guy Rosman, John J Leonard, and John W Fisher. The manhattan frame model—manhattan world inference in the space of surface normals. *IEEE transactions on pattern analysis and machine intelligence*, 40(1):235–249, 2017.

[38] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

[39] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J Davison. imap: Implicit mapping and positioning in real-time. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6229–6238, 2021.

[40] Mohammed Suhail, Carlos Esteves, Leonid Sigal, and Ameesh Makadia. Light field neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8269–8279, 2022.

[41] Jingxiang Sun, Xuan Wang, Yong Zhang, Xiaoyu Li, Qi Zhang, Yebin Liu, and Jue Wang. Fenerf: Face editing in neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7672–7682, 2022.

[42] Frank A Van den Heuvel. 3d reconstruction from a single image using geometric constraints. *ISPRS Journal of Photogrammetry and Remote Sensing*, 53(6):354–368, 1998.

[43] Carlos A Vanegas, Daniel G Aliaga, and Bedrich Benes. Building reconstruction using manhattan-world grammars. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 358–365. IEEE, 2010.

[44] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5481–5490. IEEE, 2022.

[45] Suhani Vora, Noha Radwan, Klaus Greff, Henning Meyer, Kyle Genova, Mehdi SM Sajjadi, Etienne Pot, Andrea Tagliasacchi, and Daniel Duckworth. Nesf: Neural semantic fields for generalizable semantic segmentation of 3d scenes. *arXiv preprint arXiv:2111.13260*, 2021.

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

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

[48] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[49] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021.

[50] Horst Wildenauer and Allan Hanbury. Robust camera self-calibration from monocular images of manhattan worlds. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2831–2838. IEEE, 2012.

[51] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. In *Computer Graphics Forum*, volume 41, pages 641–676. Wiley Online Library, 2022.

[52] Wenqi Yang, Guanying Chen, Chaofeng Chen, Zhenfang Chen, and Kwan-Yee K Wong. Ps-nerf: Neural inverse rendering for multi-view photometric stereo. In *European Conference on Computer Vision*, pages 266–284. Springer, 2022.

[53] Wenqi Yang, Guanying Chen, Chaofeng Chen, Zhenfang Chen, and Kwan-Yee K Wong. S$\hat{3}$-nerf: Neural reflectance field from shading and shadow under a single viewpoint. *arXiv preprint arXiv:2210.08936*, 2022.

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

[55] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5752–5761, 2021.

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

[57] Menghua Zhai, Scott Workman, and Nathan Jacobs. Detecting vanishing points using global image context in a non-manhattan world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5657–5665, 2016.

[58] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15838–15847, 2021.