

# Neural Radiance Fields for Manhattan Scenes with Unknown Manhattan Frame

Nikola Popovic<sup>1</sup>, Danda Pani Paudel<sup>1</sup>, Luc Van Gool<sup>1,2</sup>

<sup>1</sup>Computer Vision Laboratory, ETH Zurich, Switzerland

<sup>2</sup>VISICS, ESAT/PSI, KU Leuven, Belgium

{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 scene (under investigation) being Manhattan is known – with no additional information whatsoever – with an unknown Manhattan coordinate frame. Such high-level prior is then used to self-supervise the surface normals derived explicitly in the implicit neural fields. Our modeling allows us to group the derived normals, followed by exploiting 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.*

## 1. Introduction

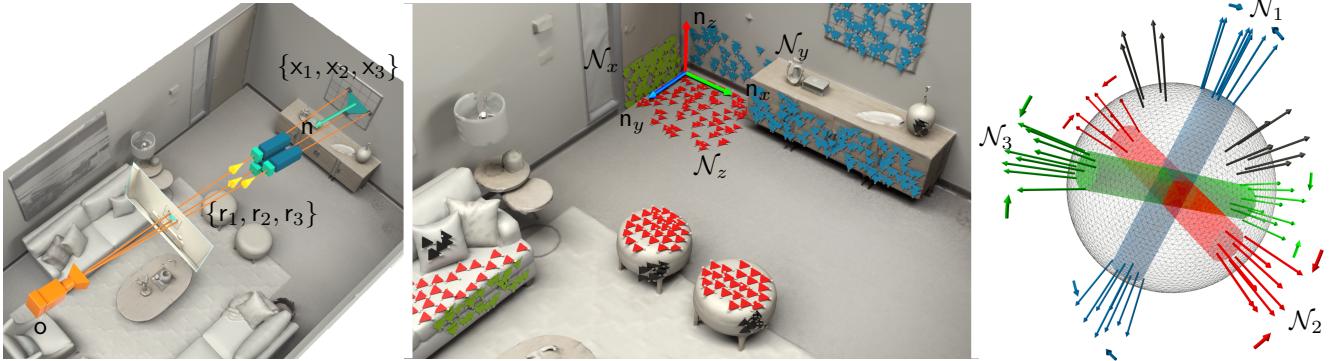
Above 80% images ever taken are estimated to involve human-made architectural structures [11]. These scenes often exhibit strong structural regularities, including flat and texture-poor surfaces in some axis-aligned Cartesian coordinate system – also known as Manhattan world [8, 12]. Paradoxically, these 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 [12, 45], camera localization [20],

self-calibration [52], and more [31, 37, 39, 60].

For calibrated multi-view cameras, 3D inversion using implicit neural representations [26, 49, 57] is becoming increasingly popular due to their remarkable performance and recent research developments for efficiency [27, 32, 54, 58]. Meanwhile, such representations are known to benefit from additional supervision in the form of depth [2, 9, 34], normal [19], semantics [15, 18, 47, 61], local-regularization [28, 46, 48], or their combinations [14, 22, 59]. In this context, two recent notable works NeurMiPs [22] and ManhattanSDF [14] demonstrate the benefit of exploiting the high-level geometric prior for structured scenes. More precisely, NeurMiPs [22] exploits the known planar patches, whereas ManhattanSDF [14] uses the known semantic regions to impose the planar geometry prior of floors and walls under the Manhattan scene assumption. In both cases, the coordinates of the planar structures are assumed partially or fully 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 Manhattan scenes, with no further assumptions. In other words, we consider that the structural and semantic information is not available, unlike NeurMiPs [22] and ManhattanSDF [14]. More importantly, we consider that the Manhattan coordinate frame is also unknown. Our assumptions (of unknown semantics, structure, and Manhattan frame) on the one hand make our setting very practical. On the other hand, the same 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 [39]. Unfortunately, recovering the Manhattan frame directly from images is not an easy task [8, 39].



**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 Manhattan prior (right).

Therefore, several methods have been developed until recently [5, 10, 13, 17, 38] to recover the Manhattan frame, relying upon the known 3D reconstruction or image primitives (eg. lines, planes). Note that the 3D representation could facilitate the Manhattan frame recovery process, thereby allowing us to exploit the Manhattan prior. On the contrary, we wish to exploit the Manhattan prior to improve the 3D representation instead. This contradiction leads us to the problem of simultaneously recovering the Manhattan frame and representing the 3D scene.

In this work, we propose a method that jointly learns the Manhattan frame and neural radiance field, from calibrated multi-views in the Manhattan scene, in an end-to-end manner using the recent efficient backbone of Instant-NGP [27]. 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 for rotation matrix whose entries are directly derived from the orthogonal clusters of normals, to align the Manhattan frame. The recovered Manhattan frame 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 both in discovering the Manhattan frame and 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 two realistic 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.* [26], implicit neural representation of 3D scenes have advanced in various fronts. These fronts include representation [42, 57], generalization [21, 24, 30, 36, 51], generation [7, 29], and efficient methods [27, 32, 58]. We rely on a recent method Instant-NGP [27] 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 [9, 35, 41], semantics [15, 18, 43, 61], normal [55, 56], and their combinations [14, 22, 59] are shown to be beneficial on improving the 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 [59] or recovered from the structure-from-motion (SfM) reconstructions [9, 22]. We argue that such labels 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* [3], Manhattan structure reasoning is done directly on/from images by detecting the so-called the vanishing-points (VPs) [25,44]. In fact, the problem of detecting three orthogonal VPs is equivalent to finding the MF in 3D for the calibrated multi-view setting [4]. Note that the knowledge of Manhattan structure has been used in several computer vision works [12, 20, 31, 37, 39, 45, 52, 60]. When unknown, most methods implicitly or explicitly estimate the MF to leverage the Manhattan scene prior. In [39], 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 [39], using the surface normals derived explicitly from the implicit neural fields.

**Scene structure prior:** Two notable works, NeurMiPs [22] and ManhattanSDF [14] demonstrate clear benefits of using scene structure prior. In this regard, NeurMiPs assumes that the SFM reconstructions with their planar decomposition are available. Scene structure prior in the form of a “union of planes” is then exploited. Note that SFM is not necessary to obtain the calibrated cameras required for the implicit neural representation of 3D. Therefore, SfM reconstruction is not always at disposal, besides it requiring additional computation and being poor in low-texture areas. Another 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. In this work, we make no assumption about the available labels. Instead, we *perform self-supervision of depth and normals* merely using the Manhattan prior. Our self-supervision seeks the Manhattan world in the normals derived from the depth map obtained using volumetric rendering.

### 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  $\mathcal{E} = \{\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z\}$ . Without loss of generality, let the axes’ coordinates be  $\mathbf{e}_x = [1, 0, 0]^\top$ ,  $\mathbf{e}_y = [0, 1, 0]^\top$ , and  $\mathbf{e}_z = [0, 0, 1]^\top$ . Note that these axes align with the normals of the respective scene building planes, in the WF. Let  $\mathcal{N} = \{\mathbf{n}_i\}_{i=1}^m$  be a set of 3D normals of all the scene planes. Then, the rotation  $R$ , from world to Manhattan frame, aligns the normals  $\mathbf{n} \in \mathcal{N}$  to the Manhattan axes  $\mathbf{e} \in \mathcal{E}$ , i.e.,  $R\mathbf{n}_i \in \mathcal{E}, \forall \mathbf{n}_i \in \mathcal{N}$ .

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  $\{\mathbf{n}_x, \mathbf{n}_y, \mathbf{n}_z\}$  from the corresponding orthogonal subsets, we aim to establish the condition  $[\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z] = R[\mathbf{n}_x, \mathbf{n}_y, \mathbf{n}_z]$ . Note that the assignment condition results into  $R = [\mathbf{n}_x, \mathbf{n}_y, \mathbf{n}_z]^\top$ . 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 [39]. Alternatively, one can also perform the robust estimation of the orthogonal subsets of  $\mathcal{N}$ , followed by solving  $[\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z] = R[\mathbf{n}_x, \mathbf{n}_y, \mathbf{n}_z], \forall \mathbf{n}_x \in \mathcal{N}_x, \mathbf{n}_y \in \mathcal{N}_y, \mathbf{n}_z \in \mathcal{N}_z$ , 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  $\mathbf{o}$  is obtained using the volumetric rendering of the implicit neural radiance field, repre-

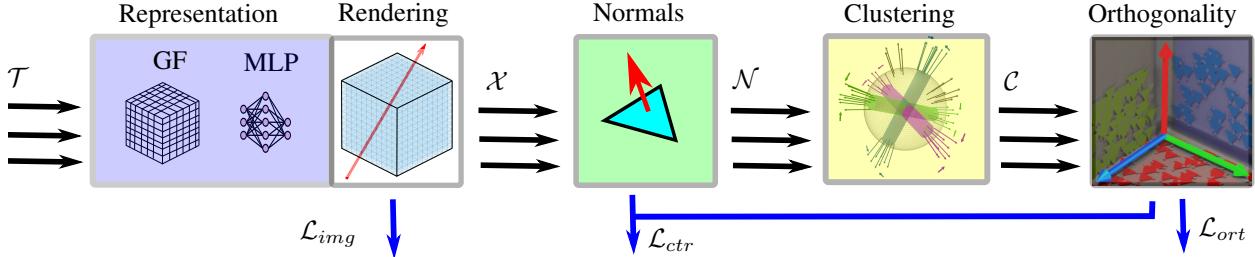


Figure 2. **The complete pipeline of our method.** We use grid features (GF) and a MLP to represent the radiance field. The explicit normals are derived using the depths obtained from volume rendering. The Manhattan scene prior is exploited by clustering the estimated normals to enforce their orthogonality.

sented by a neural network parameterized by  $\theta$  as in [26]. Then, the location of the 3D surface point is given by,  $\mathbf{x} = \mathbf{o} + d\mathbf{r}$ . We process triplets of three neighboring rays. Let  $\mathcal{T} = \{\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3\}$  be such a triplet, whose corresponding surface points are given by  $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ . Now, for the ray triplet  $\mathcal{T}$ , we obtain the explicit surface normal  $\mathbf{n}$  with the help of the point triplet  $\mathcal{X}$ , using the following mapping and algebraic operation,

$$\mathcal{T} \rightarrow \mathcal{X} \rightarrow \mathbf{v} = (\mathbf{x}_1 - \mathbf{x}_2) \times (\mathbf{x}_1 - \mathbf{x}_3) \rightarrow \mathbf{n} = \frac{\text{sign}(\mathbf{o}^\top \mathbf{v}) \mathbf{v}}{\|\mathbf{v}\|}. \quad (1)$$

Note that  $\mathbf{v}$  is a vector orthogonal to the plane passing through 3D points in triplet  $\mathcal{T}$ . We obtain the oriented normal  $\mathbf{n}$  by normalizing  $\mathbf{v}$  and correcting its sign by ensuring that the camera center  $\mathbf{o}$  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 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  $\mathcal{N} = \{\mathbf{n}_i\}_{i=1}^m$ , 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  $\mathcal{N} = \{\mathbf{n}_i\}_{i=1}^m$ . 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 well-known  $k$ -means clustering algorithm [6]. 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  $\mathcal{E} = \{\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z\}$  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  $\mathcal{U} = \{(\mathcal{C}_i, \mathbf{c}_i)\}_{i=1}^k$ , where the centroids are computed by taking the average across the corresponding cluster such that  $\mathbf{c}_i = \frac{1}{|\mathcal{C}_i|} \sum_{\mathbf{n} \in \mathcal{C}_i} \mathbf{n}$ , followed by normalizing to a unit vector  $\mathbf{c}_i = \frac{\mathbf{c}_i}{\|\mathbf{c}_i\|}$ . Leveraging the Manhattan prior through the assumption that Manhattan surfaces dominate the scene, we pick  $\mathbf{n}_1 = \mathbf{c}_f$ , 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  $\mathbf{n}_2 = \mathbf{c}_s$  and  $\mathbf{n}_3 = \mathbf{c}_t$  as a solution of the optimization problem  $s, t = \arg \min_{i,j} |\mathbf{c}_i^\top \mathbf{n}_1| + |\mathbf{n}_1^\top \mathbf{c}_j| + |\mathbf{c}_i^\top \mathbf{c}_j|$ . 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  $\{\mathbf{n}_1, \mathbf{n}_2, \mathbf{n}_3\}$  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 closest 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.

---

**Algorithm 1** ( $\mathcal{N}_1, \mathcal{N}_2, \mathcal{N}_3, R$ ) = 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  $(\mathcal{C}, c) \rightarrow (\mathcal{N}_1, n_1)$ .
  3. Find  $\mathcal{C}_s \in \mathcal{U}, \mathcal{C}_t \in \mathcal{U}$  minimizing  $|\mathbf{c}_s^\top \mathbf{n}_1| + |\mathbf{n}_1^\top \mathbf{c}_t| + |\mathbf{c}_s^\top \mathbf{c}_t|$ .
  4. Assign  $(\mathcal{C}_s, \mathbf{c}_s) \rightarrow (\mathcal{N}_2, \mathbf{n}_2), (\mathcal{C}_t, \mathbf{c}_t) \rightarrow (\mathcal{N}_3, \mathbf{n}_3)$ .
  5.  $|\mathbf{e}_z^\top \mathbf{n}_1| \leq \frac{1}{\sqrt{2}} |\mathbf{n}_1| \rightarrow \mathbf{n}_z : (|\mathbf{e}_y^\top \mathbf{n}_1| \leq \frac{1}{\sqrt{2}} |\mathbf{n}_1| \rightarrow \mathbf{n}_y : \mathbf{n}_1 \rightarrow \mathbf{n}_x)$ .
  6. Perform the remaining  $\{\mathbf{n}_1, \mathbf{n}_2, \mathbf{n}_3\} \rightarrow \{\mathbf{n}_x, \mathbf{n}_y, \mathbf{n}_z\}$  as in step 5 to the closest corresponding canonical axes.
  7. Return  $\mathcal{N}_1, \mathcal{N}_2, \mathcal{N}_3$ , and  $R = [\mathbf{n}_x, \mathbf{n}_y, \mathbf{n}_z]^\top$ .
- 

### 4.3. Self-supervision by Manhattan Prior

The supervision of the implicit neural radiance field, parameterized by  $\theta$ , is achieved through jointly optimizing three loss terms. The first one is the photometric loss computed as follows,

$$\mathcal{L}_{img} = \frac{1}{|\mathcal{R}|} \sum_{\mathbf{r} \in \mathcal{R}} \|C_\theta(\mathbf{r}) - C(\mathbf{r})\|_2^2, \quad (2)$$

where  $C(\mathbf{r})$  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 [26].

**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  $\{(\mathcal{N}_i, \mathbf{n}_i)\}_{i=1}^3$  we use the following two losses,

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

$$\mathcal{L}_{ort} = \frac{1}{3} (|\mathbf{n}_1^\top \mathbf{n}_2| + |\mathbf{n}_1^\top \mathbf{n}_3| + |\mathbf{n}_2^\top \mathbf{n}_3|). \quad (4)$$

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  $\theta$  is then given by,

$$\mathcal{L} = \mathcal{L}_{img} + \lambda_{ctr} \mathcal{L}_{ctr} + \lambda_{ort} \mathcal{L}_{ort}, \quad (5)$$

where  $\lambda_{ctr}$  and  $\lambda_{ort}$  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** [27] (**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.

**ManhattanSDF** [14]: This method exploits the Manhattan prior by supervising the explicit surface normals of the floors to align with the known  $\mathbf{n}_z$ , as well as by supervising normals of the walls to align with two orthogonal axes which are also orthogonal to  $\mathbf{n}_z$ . To achieve this, apart from the RGB images, this method relies on knowing the wall & floor semantics, as well as knowing the exact floor axis  $\mathbf{n}_z$ . We implemented it on top of the InstantNGP backbone.

**ManhattanSDF** [14] (**MF known**): We modify the ManhattanSDF by assuming that the MF is fully known. We do so by choosing  $\mathbf{n}_f$  and  $\mathbf{n}_w$  from equations (8) and (9) in [14] to be the closest MF axis (or a closer opposite counterpart).

**Ours:** We implement our proposed method from Section 4 on top of the InstantNGP backbone.

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

**Other SotA:** We also compare our method to NeurMiPs [22], by performing experiments directly in their settings. We also compare against the same SotA methods that NeurMiPs compared to, namely PlenOctrees [58], KiloNeRF [32], NeX [53] and NeRF [26]. In additional experiments, we also compare to SemanticNeRF [61], which we implement on top of the InstantNGP backbone.

**Metrics:** To quantitatively evaluate novel view rendering, we use peak signal-to-noise ratio (PSNR) and the structural similarity index (SSIM) [50]. 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 the RGB images, ManhattanNerf [14] requires the knowledge of wall, floor semantics and the exact floor axis  $\mathbf{n}_z$ , whereas NeurMiPs [22] require

Table 1. **Experiments on Hypersim.** We observe that our method consistently outperforms the baseline. Also, our method is slightly better than ManhattanSDF, but on par when ManhattanSDF knows the full MF. This is very interesting since, unlike ManhattanSDF, 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 or not.

		PSNR↑	SSIM↑	Normals° ↓	Pitch° ↓	Roll° ↓	Yaw° ↓	D-MAE↓	D-RMSE↓
Scenes A	InstantNGP [27] (baseline)	25.86	0.871	57.12	6.18	6.46	19.25	0.064	0.102
	ManhattanSDF [14]	26.53	0.868	39.51	1.21	1.26	7.26	0.051	0.086
	ManhattanSDF [14] (MF known)	27.22	0.871	46.20	0.67	0.66	1.52	0.050	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 [27] (baseline)	20.75	0.811	60.12	6.06	7.92	15.87	0.105	0.151
	ManhattanSDF [14]	21.87	0.826	50.50	2.55	2.06	11.69	0.079	0.121
	ManhattanSDF [14] (MF known)	22.39	0.826	51.08	0.78	2.08	2.17	0.073	0.115
	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 [27] (baseline)	17.79	0.740	64.29	7.45	4.55	15.14	0.130	0.174
	ManhattanSDF [14]	18.33	0.770	56.08	3.20	3.41	10.25	0.103	0.147
	ManhattanSDF [14] (MF known)	19.65	0.779	56.24	1.98	1.35	4.92	0.080	0.123
	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 [27] (baseline)	20.47	0.783	61.34	6.56	6.99	21.75	0.104	0.146
	ManhattanSDF [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

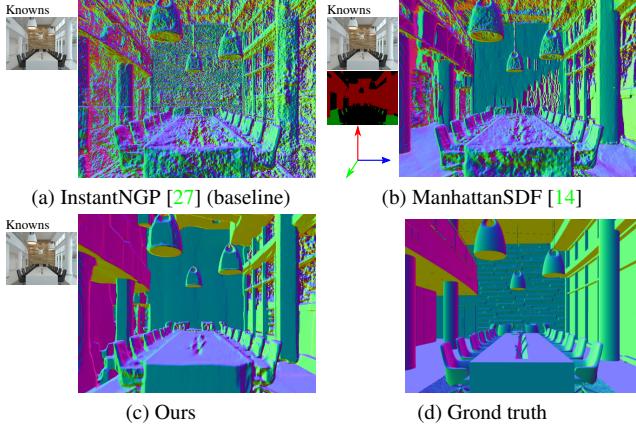


Figure 3. **Leveraging the Manhattan prior.** Our method leverages the presence of many Manhattan objects and areas in the scene in a self-supervised fashion, which facilitates imposing geometrical constraints during learning. In contrast, ManhattanSDF leverages the Manhattan prior through labels of walls & floors.

planar patches extracted from SfM reconstruction. In contrast, our method only requires RGB images and the assumed Manhattan prior to hold true. Therefore, we compare against these methods to get a better insight into what can be achieved without leveraging additional information.

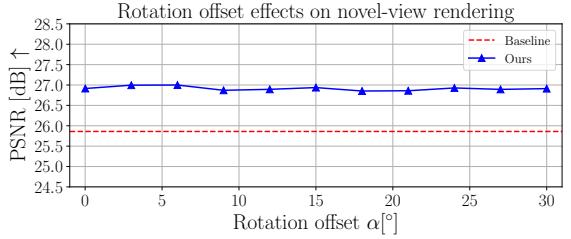
## 5.2. Datasets

We experiment on two different datasets to demonstrate the versatility and effectiveness of our approach.

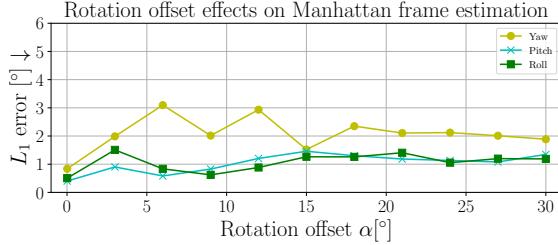
**Hypersim** [33] is a photorealistic synthetic dataset consisting of indoor scenes. It was created by leveraging a large repository of scenes created by professional artists, with 461 indoor scenes in total. This dataset is geometry-rich,

containing a lot of details and lighting sources. Each scene has one or more camera trajectories available, where each trajectory has up to 100 rendered views 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 [27] 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 often the test split contains views that were partially or completely unobserved during training.

**Replica** [40] is a synthetic dataset featuring a diverse set of 18 high-fidelity indoor scenes. Each scene is equipped with high-quality geometry and photorealistic textures, allowing one to render realistic images from arbitrary camera poses. We use two different versions of the Replica dataset. In order to directly evaluate our method in the settings of [22], we use **Replica-1** which contains 7 scenes rendered in [22]. The authors focused on challenging novel view synthesis scenarios such as extreme view extrapolation. Each scene contains 50 training images and 100 test images in 512×512 resolution, where the camera poses were sampled randomly within a pre-defined range. Replica-1 also contains dense point clouds extracted through SfM, which are necessary for [22] to extract planar patches. Unfortunately, Replica-



(a) Novel-view rendering quality remains robust to the angular offset  $\alpha$ .



(b) MF estimation remains robust to the angular offset  $\alpha$ .

Figure 4. **Finding the MF.** We test the robustness of our method by changing offsets between the MF and the WF.

1 does not have rendered semantic information required for the ManhattanSDF [14]. Therefore, we use Replica with semantics, named **Replica-2** here, obtained by taking 5 scenes rendered in [61]. Each scene contains 900 rendered views generated from random 6-DOF trajectories similar to handheld camera motions in 640x480 resolution. We select 75 views for training and 75 for testing in each scene. We note that even though two scenes overlap in Replica-1 and Replica-2, all scenes drastically differ in rendered appearance, texture, and camera trajectories.

We also note that Hypersim is visually more realistic, geometry-rich, and challenging. This can be objectively observed by considering that the InstantNGP [27] baseline achieves 20.47 PSNR on Hypersim scenes while achieving 28.36 and 34.30 PSNR on Replica-1&2, respectively.

### 5.3. Results

**Experiments on Hypersim:** We summarize experiments performed on the Hypersim dataset in Table 1. Our proposed method achieves clear improvements in comparison to the InstantNGP baseline, in terms of PSNR, surface normals, recovered MF, and depth. This is consistent across different scene difficulties, namely splits A, B & C. Furthermore, our method performs slightly better than ManhattanSDF when estimating surface normals and recovering the MF, but the performance is however on par when ManhattanSDF knows the MF. This is very interesting since, unlike ManhattanSDF, we do not use any additional labels during training, other than the ground truth RGB. The ManhattanSDF has access to one or all of the MF axis, as well as wall & floor semantics, which directly implies to having sparse surface normals ground truth. On the other hand, there are lots of Manhattan objects and areas in realistic

Table 2. **Experiments on Replica-1.** Our method performs relatively close to NeurMiPs, even though NeurMiPs relies on additional geometric information during training.

		PSNR↑	SSIM↑
Grid based	PlenOctrees [58]	27.72	0.872
	KiloNeRF [32]	29.37	0.904
	InstantNGP [27]	28.36	0.913
	Ours	29.70	0.911
Deep network based	NeX [53]	24.76	0.832
	NeRF [26]	30.12	0.901
	NeurMiPs [22]	30.80	0.900

scenes such as in Hypersim, other than the walls and floors. Our self-supervised method leverages the presence of many such objects and areas, in order 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 or unknown, since our method recovers the MF in the clustering step automatically, and therefore similar performance is achieved in both cases. Finally, we show large-scale experiments on a subset of 194 scenes, which lead to the same conclusions. The semantic loss of the ManhattanSDF was causing convergence issues on many difficult scenes, which we partially alleviated with class weighting and label smoothing. We report all results on 194 scenes where ManhattanSDF converged. Additional results on all 435 scenes are provided in the supplementary.

**Finding the MF:** In Hypersim, MF is equivalent to the world frame. We test the robustness of our proposed method for finding the Manhattan frame, by introducing a simultaneous offset  $\alpha$  in the yaw, pitch, and roll of the WF. The experiments are performed on Hypersim A and can be found in Figure 4. In Figure 4a we observe that the novel-view rendering quality remains largely robust to the angular offset  $\alpha$ . We note that we increase the scene bounding box by the same factor for all experiments in Figure 4, to make sure that objects remain within the voxel grid for maximal  $\alpha$ . This slightly decreases the resolution of grid element representations, so PSNR is slightly lower than in Table 1. Furthermore, in Figure 4b we see that the MF estimation remains robust with respect to the angular offset  $\alpha$ .

**Experiments on Replica-1:** We compare our approach to NeurMiPs [22] directly in one of their experimental settings, namely Replica-1, and summarize the results in Table 2. This dataset contains many unobserved views

Table 3. **Experiments on Replica-2.** Our method is superior to the baseline, and it shows similar performance as ManhattanSDF, which leverages additional geometrical labels during training.

	PSNR↑	SSIM↑	Depth↓-MAE
InstantNGP [27] (baseline)	34.30	0.944	0.022
Semantic-NeRF [61]	34.08	0.938	0.014
ManhattanSDF [14]	35.24	0.944	0.008
Ours	35.13	0.944	0.011

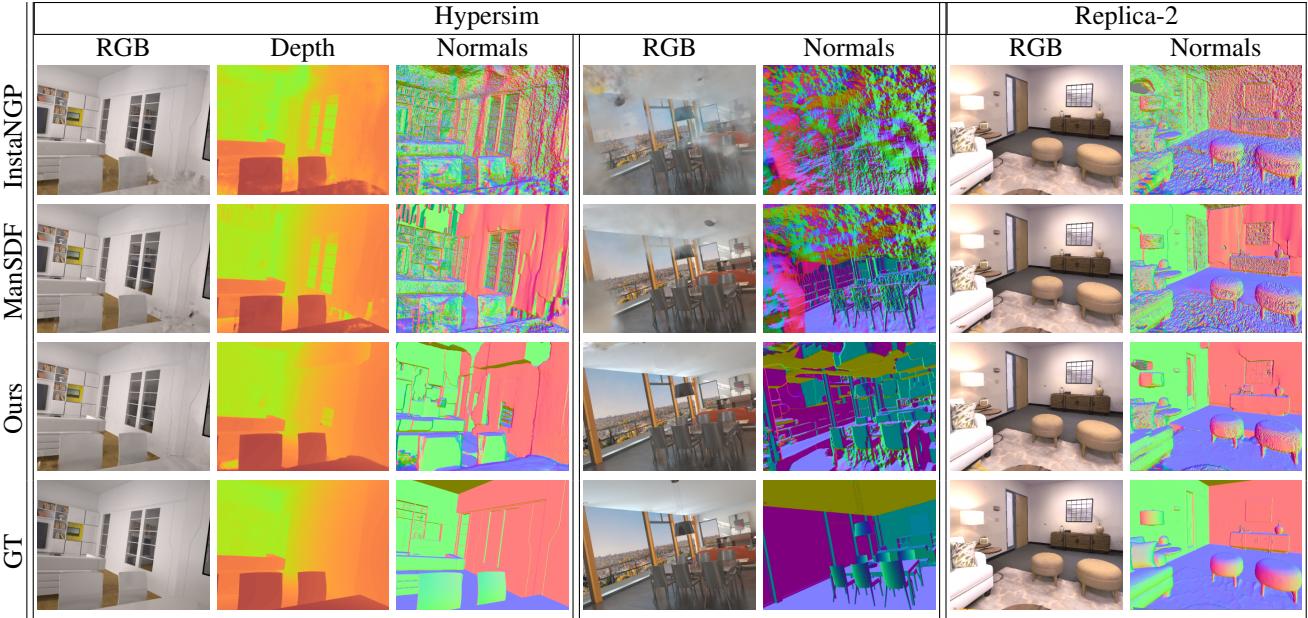


Figure 5. **Qualitative results.** Our method leverages many Manhattan objects and surfaces in the scene, which improves the implicit geometrical representation compared to the baseline. Unlike ManhattanSDF, our method relies on many cues other than the walls & floors, which leads to a better representation of some objects (e.g. the white table in the bottom right corner of the example from columns 1-3).

in the test set, which are problematic for grid-based approaches [22]. Despite this, our method performs relatively close to NeurMiPs (-1.1 PSNR, +0.011 SSIM). This is again very interesting, since NeurMiPs relies on additional geometrical information during training, in the form of known planar patches extracted from SfM. Our method however only leverages the Manhattan prior in a self-supervised fashion. We also note that our method performs best among the grid-based approaches.

**Experiments on Replica-2.** In order to gain more insight, we perform additional experiments on Replica-2, which are presented in Table 3. We observe that our method is still superior to the baseline. However, ManhattanSDF performs only slightly better. Again, we achieve almost the same performance as a method that leverages additional geometrical labels during training. We also note that Replica-2 has noticeably more walls & floors, and less of other objects, compared to the more complex Hypersim dataset.

**Ablation study:** We report our ablation study in Table 4.

Table 4. **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

Both of our proposed losses contribute to the overall performance. Furthermore, turning on  $\mathcal{L}_{ort}$  and  $\mathcal{L}_{centr}$  after 500 steps, and linearly increasing their weight also helps slightly. Please refer to the supplementary for more details.

**Qualitative results:** We depict the visual results of the discussed experiments in Figure 5. 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 ManhattanSDF, 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 from columns 1-3. Moreover, we observe that our method is able to cope with difficult scenes and views, where other methods struggle (columns 4,5). For more qualitative results, please refer to the supplementary material.

## 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 together into three orthogonal directions. Our experiments on two 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.

## A. Appendix

This appendix provides additional details, which complement the main paper. We first give additional details about our implementation and experimental setup, which are contained in Section A.1. In Section A.2, we complement the results from the main paper by showing additional comparisons and studies. Furthermore, Section A.3 comments about the limitations and failures of our approach. Finally, we provide additional qualitative examples in Section A.4.

### A.1. Implementation Details

In this paragraph we describe the experimental setup when using the InstantNGP backbone. We use a hash table of size  $2^{19}$  containing 16 levels with 2 features per level, maximum grid resolution of 2048, and an occupancy grid of resolution 128. We jointly train the neural network weights and the hash table entries by applying Adam with  $\epsilon = 10^{-15}$  and a learning rate of  $10^{-2}$ . We also apply  $\mathcal{L}_2$  regularization with a factor of  $10^{-6}$ , but only on neural network weights. These choices are based on suggestions in [27]. Also, for the sake of efficiency, we update the density grid after every 16 steps similarly to the procedure described in [27]. To further stabilize the training we also use the opacity regularization [23] with a factor of  $10^{-3}$ , as well as gradient norm clipping with a factor of 0.05. We also use a cosine annealing learning rate scheduler and perform each training on 30k iterations with batch size 8190.

When it comes to our proposed method we set the loss weights  $\lambda_{ctr} = \lambda_{ort}$  to  $2 \cdot 10^{-3}$  in the case of Hypsim,  $10^{-4}$  in the case of Replica-1, and  $5 \cdot 10^{-4}$  in the case of Replica-2. We turn on  $\mathcal{L}_{ort}$  and  $\mathcal{L}_{ctr}$  after 500 steps, and linearly increase their  $\lambda$  weights to the specified values over the next 2500 steps. Also, when using methods which rely on explicit surface normals, we randomly sample rays for one-third of every batch size and select their left and upper pixel neighbor to form a triplet to facilitate obtaining these explicit normals. We call this a triplet triangle. Every triangle in the batch is sampled randomly from a set of all possible triangle triplets of all available images.

For k-means clustering, we use  $k = 20$  clusters when processing training batches. However, in order to estimate

the Manhattan frame for all methods after the training ends, we cluster the normals from the whole test set into  $k = 30$  clusters. In addition, during both training and testing, we merge the three selected orthogonal clusters with their opposites. This is achieved by comparing the similarity of a selected cluster centroid  $n_i$  and the opposite vector of every other cluster centroid  $(-c_j)$ . In the case  $|n_i^T(-c_j)| > 1 - t$  holds true, all cluster elements of  $\mathcal{C}_j$  are multiplied by  $-1$  and added to  $\mathcal{N}_i$ . Also, if any cluster centroid  $c_j$  is close to one of the selected centroids  $n_i$  ( $|n_i^T c_j| > 1 - t$ ), elements from  $\mathcal{C}_j$  are added to  $\mathcal{N}_i$ .

During the implementation of ManhattanSDF [14], we initially had stability issues with the semantic segmentation cross-entropy loss. In order to alleviate these issues, we used label smoothing with a factor of 0.1, as well as a class weight of 0.3 for the background class. Also, in the Hypsim dataset, the scenes are much richer in content, and there are less wall & floor labels compared to Replica. Therefore we merged the floor class with the floor mat class, and we also merged the window class with the wall class. We also observed in Hypsim that in a small number of scenes the ceiling is labeled as wall, or that there are no wall & floor labels since they are all labeled as void.

Also, when it comes to comparing baselines and different methods on different datasets, we always perform a thorough search over additional hyperparameters before reporting results. On Hypsim the weight for the semantic loss was  $10^{-1}$ , while the weight for surface normals loss was  $5 \cdot 10^{-4}$  in the regular case, and  $10^{-2}$  in the case when the Manhattan frame is known. For Replica-2, the weight for the semantic loss was  $4 \cdot 10^{-2}$ , while the weight for surface normals loss was again  $5 \cdot 10^{-4}$ .

### A.2. Further Analysis

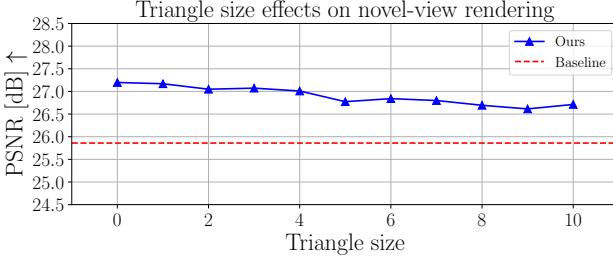
In this section, we perform further analyses of the experiments presented in Section 5 of the main paper.

**Results in 435 scenes of Hypsim:** In Table 5 we present results on all Hypsim scenes. Our proposed method performs better than the baseline with respect to all observed metrics. We were not able to evaluate ManhattanSDF on all scenes, because experiments on a large portion of the scenes had convergence issues related to the specific loss function.

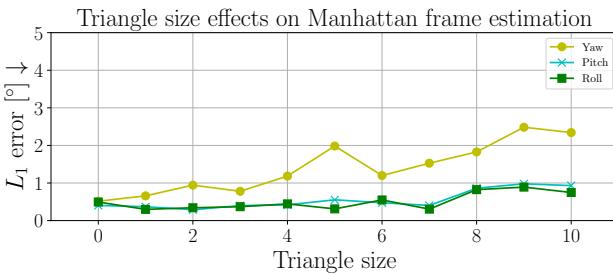
**Triplet triangle size:** As previously mentioned, when using methods that rely on explicit surface normals, we randomly sample rays for one-third of every batch size and select their left and upper pixel neighbor to form a triplet triangle to facilitate obtaining explicit normals. We call this a triangle of size 0, since the immediate left and upper neighbor are taken to form a triplet. Similarly, a triangle of size  $k$  is when there is a  $k$  pixel gap between the selected pixel and its left and upper triplet pair. In Figure 6 we show results of our method for different triangle sizes. As the triangle size increases, the performance of our method slightly decreases

Table 5. **Experiments on all Hypersim scenes.** We observe that our method outperforms the baseline with respect to all observed metrics.

		PSNR↑	SSIM↑	Normals° ↓	Pitch° ↓	Roll° ↓	Yaw° ↓	D-MAE↓	D-RMSE↓
435 scenes	InstantNGP [27] (baseline)	19.17	0.729	61.88	7.12	7.08	21.99	0.119	0.162
	Ours	20.17	0.737	54.93	3.15	3.05	8.17	0.100	0.142



(a) Triangle size choice effects on novel-view rendering.



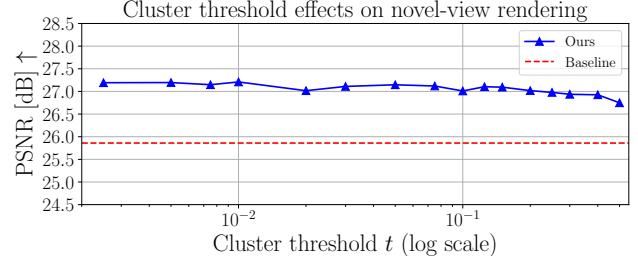
(b) Triangle size choice effects on MF estimation.

Figure 6. **Triplet triangle size choice effects on Hypersim-A.** The performance of our method slightly decreases with the triangle size increase. This is due to a bigger probability that the triplet will not lie on a planar surface segment, and thus the estimated normals will contain certain error.

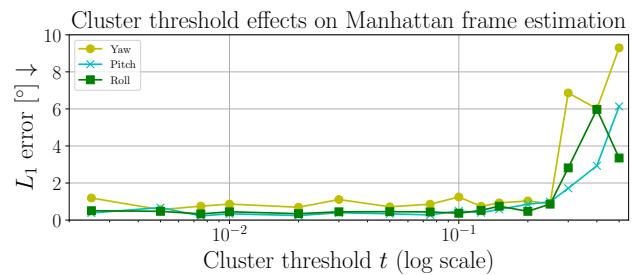
in terms of novel-view rendering as well as MF estimation. With larger triangle sizes, there is a bigger probability that the triplet will not lie on a planar surface segment and thus normals estimated with a planar assumption will contain a certain degree of error. Nevertheless, this study shows that whenever one is certain about the bigger planar regions, our method of explicit normal computation can successfully be applied to potentially gain both efficiency and performance. The further study in this regard however is out of the scope of this work, which remains as our future work.

**Cluster threshold:** We analyze the sensitivity of our proposed method to the threshold parameter  $t$ , used to merge selected orthogonal clusters with their opposites and their nearby clusters. In Figure 7, we see that our algorithm is not very sensitive to different threshold  $t$  values below 0.1.

**Ray sampling:** As previously mentioned, when using methods which rely on explicit surface normals, we randomly sample pixel triplet triangles from a set of all possible triangles in all available images. However, when using the InstantNGP baseline, we randomly sample rays from the set of all available rays in all available images. The number



(a) Cluster threshold choice effects on novel-view rendering.



(b) Cluster threshold choice effects on estimating the MF.

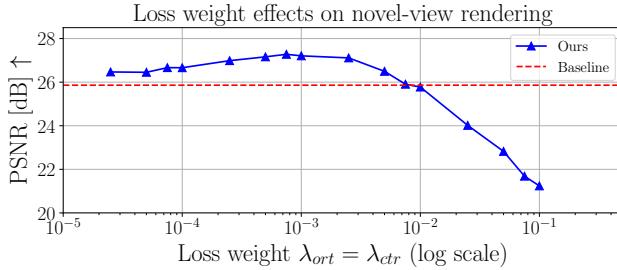
Figure 7. **Cluster threshold  $t$  choice effects on Hypersim-A.** Our method is not very sensitive to threshold values below 0.1.

of selected rays is always the same in total as the specified batch size, for all methods. In Table 6, we see that there is no significant difference if we sample triplet triangles during the baseline training instead of just randomly sampling rays. Therefore the improvements in our approach come from the proposed algorithm, and not the slightly different way of batch sampling.

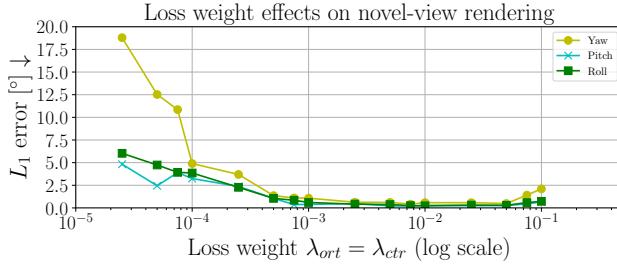
**Sensitivity in hyperparameters selection:** We further analyse the sensitivity of the loss weights  $\lambda_{ort} = \lambda_{ctr}$  multiplying our proposed loss terms, to the overall performance. The quantitative results are presented in Figure 8. Very low  $\lambda$  values lead to a bad MF estimation, whereas very high  $\lambda$  values lead to bad novel-view render-

Table 6. **Ray sampling strategy effects Hypersim-A.** There is no significant difference for the InstantNGP baseline between sampling random triplet triangles or random rays during training.

	PSNR↑	D↓-MAE	Norm. ° ↓	Rot. ° ↓
Random rays (same image)	25.72	0.072	60.11	12.36
Random rays (random images)	25.86	0.064	57.12	10.63
Random triangles (random images)	26.16	0.061	56.72	10.25



(a) Loss weight effects on novel-view rendering.



(b) Loss weight effects on estimating the MF.

Figure 8. **Loss weight  $\lambda_{ort} = \lambda_{ctr}$  choice effects on Hypersim-A.** Very low weights lead to bad MF estimation, whereas very high weights lead to bad novel-view rendering.

ing. A good trade-off is achieved around the chosen value of  $2 \cdot 10^{-3}$ . In addition, visual results are presented in Figure 10, where we can observe that very low  $\lambda$  produces noisy explicit surface normals, whereas very high  $\lambda$  makes all details appear “blocky” in normals. Moreover, this analysis has been performed statistically, on 20 scenes from the Hypersim-A split using the same hyperparameters. We observe that the scene-specific tuning of these hyperparameters could further improve the performance. These scene-specific hyperparameters differ only slightly from the ones selected for the whole set of scenes. We however, do not perform scene-specific tuning, demonstrating the generalizability of our method across the diverse scenes of Hypersim.

**Runtime:** We further analyze the run time of our proposed method. We use Python and the PyTorch Deep Learning library. The InstantNGP baseline implementation is adapted from [1], where some functionalities (e.g. volume rendering) are efficiently implemented directly in CUDA code. In addition to the baseline, our proposed approach computes explicit surface normals for every batch, followed by k-means clustering (implemented using the FI-ASS library [16]), and finally followed by computing two additional loss terms  $\mathcal{L}_{ctr}$  and  $\mathcal{L}_{ort}$ . The baseline needs  $22.77 \pm 2.07$  minutes on average to train on Hypersim-A, whereas our method needs  $29.06 \pm 1.64$  minutes. The inference time is the same for both methods, and it takes about half a minute on average to render the test set. The experiments were performed on a single NVIDIA GeForce RTX

2080 Ti GPU, with 11 Gb of RAM memory.

### A.3. Limitations

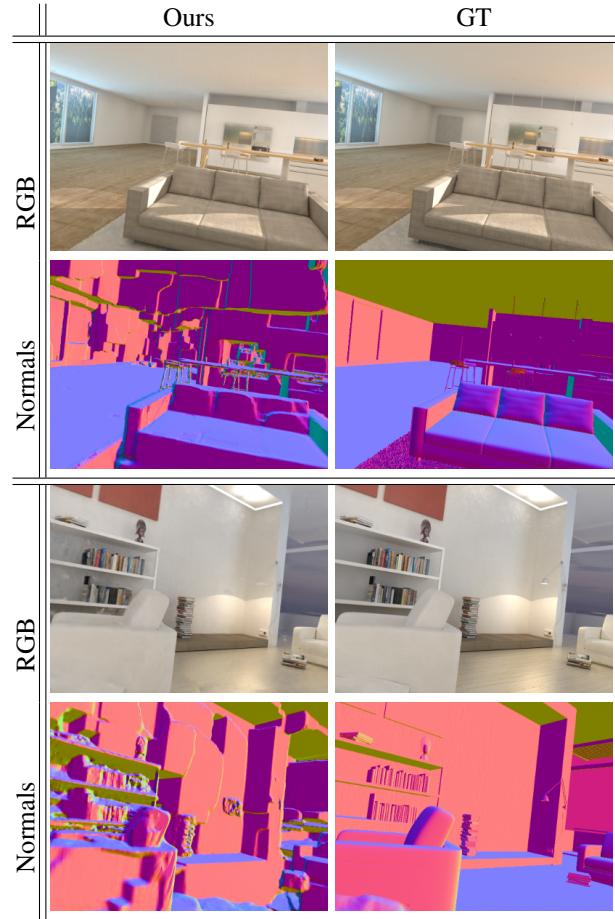


Figure 9. **Common failure cases.** In the first example, we see our method merging explicitly computed surface normals of one Manhattan direction (vertical ceiling surface) with one of the other two directions (horizontal wall surface). In the second example we see severe “blocky” artifacts in the normals. However, in both cases, there is not much trouble, nor big artifacts, when performing novel-view rendering.

While inspecting qualitative results of novel-view rendering, we observed few limitations and failure cases that occurred occasionally. Sometimes our method merges one Manhattan direction with one of the other two directions, e.g. producing same surface normals for a roof (horizontal surface) and one of the walls (vertical surface). This can be seen in the first example in Figure 9. Another phenomenon is having severe “blocky” artifacts which makes the geometry of the scene appear unnatural in the computed explicit normals, e.g. as in the second example in Figure 9. However, often in these cases, novel-view rendering does not have big problems and strong artifacts, unlike

the baseline which often struggles in similar difficult cases. This behaviour is also related to the choice of loss weights  $\lambda_{ort} = \lambda_{ctr}$ , discussed in Section A.2 of this supplementary material, as well as in Figure 10.

#### A.4. More Qualitative Results

We provide more qualitative results related to experiments from Section 5 of the main paper. Figure 11 depicts visual results from the Hypersim-A dataset, whereas Figure 12 depicts visual results from the Replica-2 dataset. We again see that 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, unlike ManhattanSDF, our method relies on many Manhattan cues other than the walls & floors. For example, different component and parts of furniture and stairs respect the Manhattan assumption, which is successfully exploited by our method.

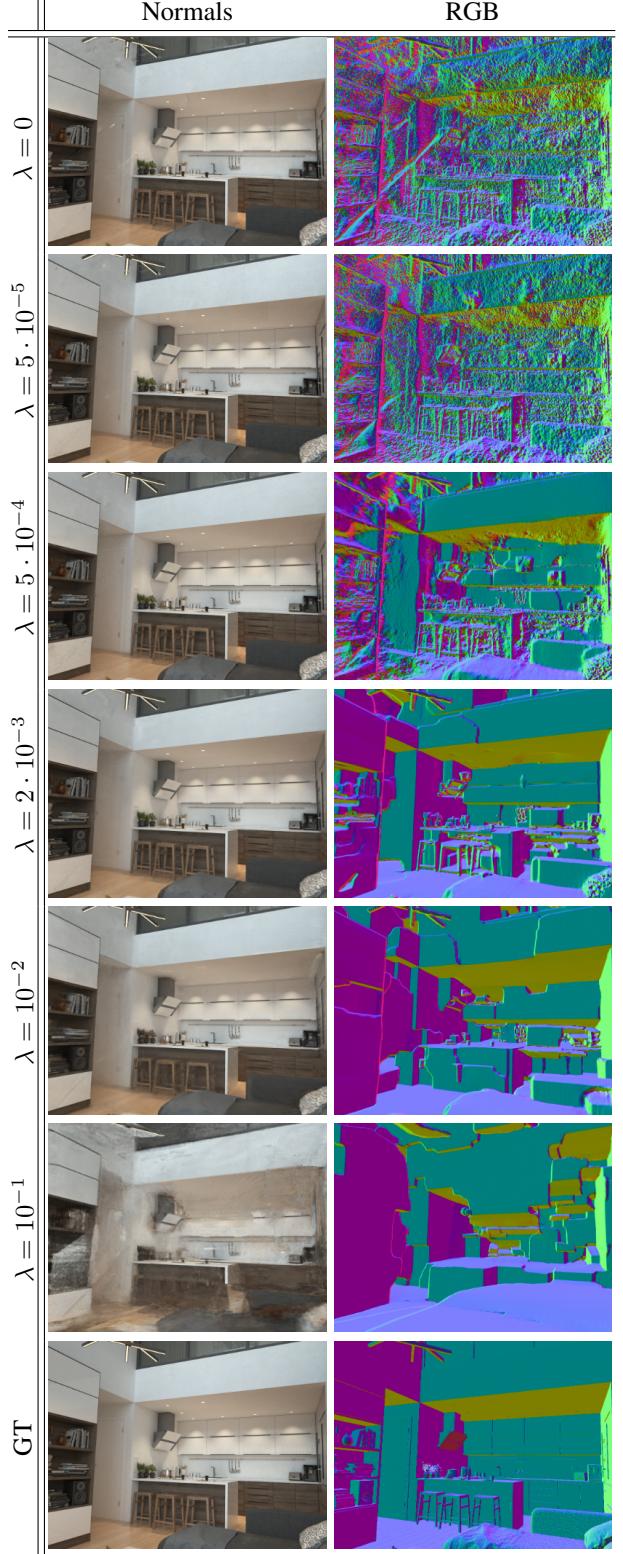


Figure 10. **Loss weight  $\lambda_{ort} = \lambda_{ctr}$  choice effects on Hypersim-A.** Very low  $\lambda$  values lead to very noisy explicit surface normals, whereas very high  $\lambda$  values lead to "blocky" artifacts in normals and bad novel-view rendering.

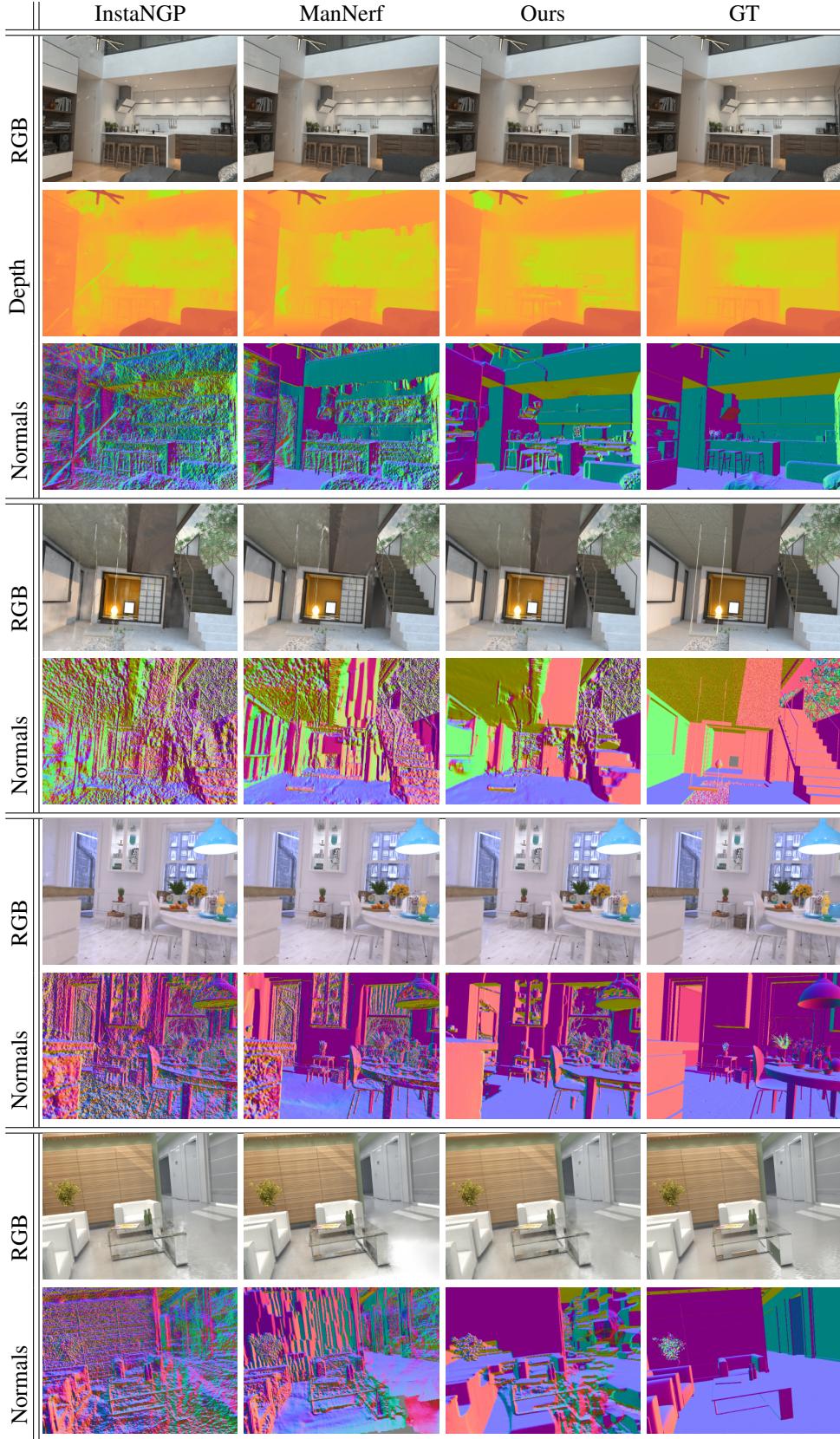


Figure 11. **Qualitative results on Hypersim-A.** Our method leverages many Manhattan objects and surfaces in the scene, which improves the implicit geometrical representation compared to the baseline. Unlike ManhattanSDF, our method relies on many Manhattan cues other than the walls & floors.

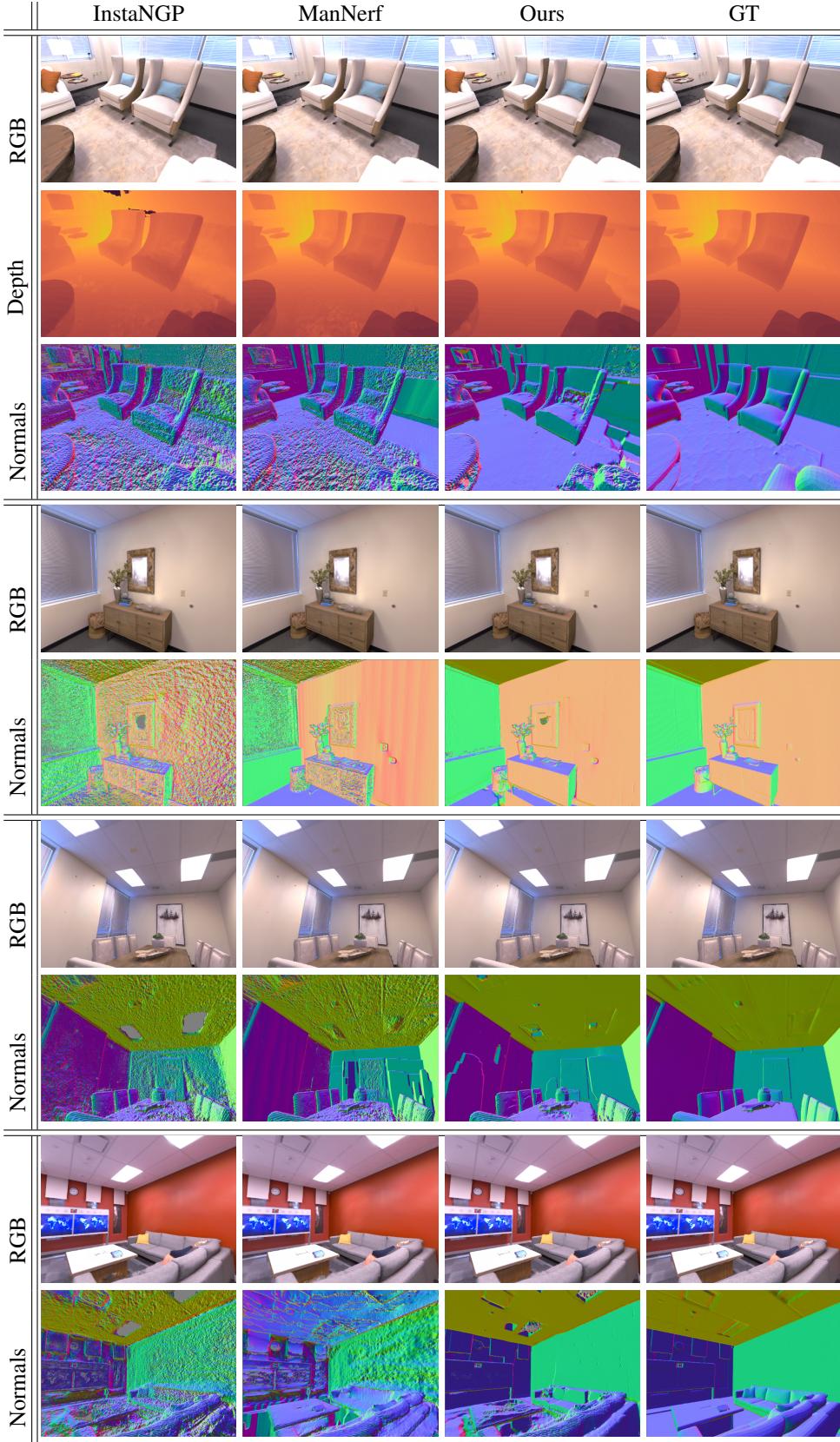


Figure 12. **Qualitative results on Replica-2.** Our method leverages many Manhattan objects and surfaces in the scene, which improves the implicit geometrical representation compared to the baseline. Unlike ManhattanSDF, our method relies on many Manhattan cues other than the walls & floors.

## References

- [1] [https://github.com/kwea123/ngp\\_pl](https://github.com/kwea123/ngp_pl). Last checked : 18.11.2022. 11
- [2] 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. 1
- [3] Stephen T Barnard. Interpreting perspective images. *Artificial intelligence*, 21(4):435–462, 1983. 3
- [4] 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. 3
- [5] 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. 2
- [6] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. 4
- [7] 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. 2
- [8] 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. 1
- [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. 1, 2
- [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. 2
- [11] Clara Fernandez-Labrador. Indoor scene understanding using non-conventional cameras. *PhD thesis*, 2020. 1
- [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. 1, 3
- [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. 2
- [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. 1, 2, 3, 5, 6, 7, 9
- [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. 1, 2
- [16] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. 11
- [17] 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. 2
- [18] 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. 1, 2
- [19] 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. 1
- [20] 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. 1, 3
- [21] 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. 2
- [22] 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. 1, 2, 3, 5, 6, 7, 8
- [23] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes. *ACM Transactions on Graphics*, 38(4):1–14, aug 2019. 9
- [24] 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. 2
- [25] 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. 3
- [26] 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>. 1, 2, 4, 5, 7
- [27] 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. 1, 2, 5, 6, 7, 9, 10
- [28] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Reg-nerf: 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. 1
- [29] 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. 2
- [30] 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. 2
- [31] 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. 1, 3
- [32] 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. 1, 2, 5, 7
- [33] 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. 6
- [34] 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. 1
- [35] 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. 2
- [36] 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. 2
- [37] Gautam Singh and Jana Kosecka. Visual loop closing using gist descriptors in manhattan world. In *ICRA omnidirectional vision workshop*, pages 4042–4047, 2010. 1, 3
- [38] 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. 2
- [39] 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. 1, 3
- [40] 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. 6
- [41] 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. 2
- [42] 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. 2
- [43] 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. 2
- [44] 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. 3
- [45] 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. 1, 3
- [46] 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. 1
- [47] 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. 1
- [48] 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. 1
- [49] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *Advances in Neural Information Processing Systems*, 34:27171–27183, 2021. 1

- [50] 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. 5
- [51] 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. 2
- [52] 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. 1, 3
- [53] Suttisak Wizadwongsu, Pakkapon Phongthawee, Jiraphon Yenphraphai, and Supasorn Suwajanakorn. Nex: Real-time view synthesis with neural basis expansion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 5, 7
- [54] 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. 1
- [55] 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. 2
- [56] Wenqi Yang, Guanying Chen, Chaofeng Chen, Zhenfang Chen, and Kwan-Yee K Wong. S<sup>3</sup>-nerf: Neural reflectance field from shading and shadow under a single viewpoint. *arXiv preprint arXiv:2210.08936*, 2022. 2
- [57] Lior Yariv, Jitao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021. 1, 2
- [58] 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. 1, 2, 5, 7
- [59] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Satler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *arXiv preprint arXiv:2206.00665*, 2022. 1, 2
- [60] 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. 1, 3
- [61] 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. 1, 2, 5, 7