

# Improving Geometric Consistency for 360-Degree Neural Radiance Fields in Indoor Scenarios

Iryna Repinetska<sup>2</sup>, Anna Hilsmann<sup>1</sup> <sup>a</sup> and Peter Eisert<sup>1,2</sup> <sup>b</sup>

<sup>1</sup>Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute, Berlin, Germany

<sup>2</sup>Department of Computer Science, Humboldt University, Berlin, Germany

**Keywords:** Novel View Synthesis, Neural Radiance Fields, Geometry Constraints, 360-Degree Indoor Dataset.


**Abstract:** Photo-realistic rendering and novel view synthesis play a crucial role in human-computer interaction tasks, from gaming to path planning. Neural Radiance Fields (NeRFs) model scenes as continuous volumetric functions and achieve remarkable rendering quality. However, NeRFs often struggle in large, low-textured areas, producing cloudy artifacts known as "floaters" that reduce scene realism, especially in indoor environments with featureless architectural surfaces like walls, ceilings, and floors. To overcome this limitation, prior work has integrated geometric constraints into the NeRF pipeline, typically leveraging depth information derived from Structure from Motion or Multi-View Stereo. Yet, conventional RGB-feature correspondence methods face challenges in accurately estimating depth in textureless regions, leading to unreliable constraints. This challenge is further complicated in 360-degree "inside-out" views, where sparse visual overlap between adjacent images further hinders depth estimation. In order to address these issues, we propose an efficient and robust method for computing dense depth priors, specifically tailored for large low-textured architectural surfaces in indoor environments. We introduce a novel depth loss function to enhance rendering quality in these challenging, low-feature regions, while complementary depth-patch regularization further refines depth consistency across other areas. Experiments with Instant-NGP on two synthetic 360-degree indoor scenes demonstrate improved visual fidelity with our method compared to standard photometric loss and Mean Squared Error depth supervision.


## 1 INTRODUCTION

Neural Radiance Fields (NeRFs) provide a novel solution to a fundamental challenge in computer vision: generating new views from a set of posed 2D images (Arandjelović and Zisserman, 2021). By modeling a scene as a continuous volumetric function and encoding it into the weights of a neural network, NeRFs achieve a remarkable balance between geometry and appearance representation (Mildenhall et al., 2021). This technology offers substantial benefits across applications such as virtual reality, augmented reality, and robotics, where high-fidelity visualization is critical.

While NeRFs produce realistic renderings across diverse settings, indoor environments with large, low-textured surfaces—such as walls, floors, and ceilings—present unique challenges. These areas

often lack distinctive visual features, which significantly hinders NeRF's ability to accurately reconstruct the scene, leading to potential inaccuracies that compromise the quality of the final render (Wang et al., 2022). This often leads to undesired artifacts in the rendered views, with one of the most common being "floaters" (Roessle et al., 2022). They appear as cloudy, erroneous, detached elements within the scene, significantly degrading the visual quality and realism of the generated views. Their occurrence is closely tied to irregularities in density distribution, stemming from inaccurate geometric estimates during color-driven optimization (Roessle et al., 2022). Imposing geometric constraints through depth supervision mitigates these issues, typically involving the comparison of rendered depth with ground truth data during the training process (Deng et al., 2022). However, acquiring accurate depth priors is an inherently challenging task, as most depth estimation methods rely on visual cues such as texture, edges, and shading to determine depth, often leading to inaccuracies in

<sup>a</sup>  <https://orcid.org/0000-0002-2086-0951>

<sup>b</sup>  <https://orcid.org/0000-0001-8378-4805>

featureless areas (Gasperini et al., 2023)—a common characteristic of many views in indoor datasets. Additionally, capturing views with a 360-degree camera further complicates the task. Since the “inside-out” viewing direction results in sparse visual overlap between adjacent images, it is harder to align features across views (Chen et al., 2023).

To overcome these challenges, we introduce an efficient method for extracting dense depth priors specifically for large planar architectural surfaces in indoor spaces, such as ceilings, walls, and floors, which are particularly susceptible to floaters. Our approach is tailored to indoor environments, requiring basic conditions that are easily met in typical settings, such as aligning the Z-axis with the floor plane normal. We assume the scene to be captured by a 360-degree camera which efficiently scans the entire rooms while being moved through the scene. Mounted on a tripod or stand, it enables straightforward estimation of the ground plane. Additionally, we assume that the room’s height is known or can be measured, which is generally true for most indoor settings. Our method is also supported by semantic segmentation information of the image data, providing class labels for wall, floor, and ceiling. Given the advanced state of current semantic segmentation techniques, numerous pre-trained models are available that can generate segmentation masks for these classes without requiring a computationally intensive training process (Chen et al., 2017), (Ronneberger et al., 2015), (Badrinarayanan et al., 2017).

Recognizing that architectural surfaces delineate the boundaries of an indoor scene, we introduce a loss function that encourages the alignment of a ray’s termination with these boundary surfaces—walls, floor, and ceiling. This function also promotes the correct distribution of volumetric densities along the ray, ensuring that the regions the ray passes through before hitting a boundary represent empty space, while density increases sharply at the boundary surfaces.

To further address flawed density distribution in other areas, we implement a patch-based depth regularization method that utilizes bilateral or joint bilateral filtering to smooth out depth inconsistencies while preserving edge information.

To evaluate our approach, we created two synthetic 360-degree indoor scenes. Rather than relying on stitched panoramic views, we propose an unconventional method that uses a series of unstitched views, facilitating precise estimation of both extrinsic and intrinsic camera parameters—critical for NeRFs pipelines—and avoiding the geometric distortions introduced by the typical stitching process. Additionally, we assume the 360-degree camera is mounted on

a movable stand, enabling efficient capture of an entire room and supporting dense depth estimation of architectural surfaces.

Our results, demonstrated on a 360-degree indoor dataset with Instant-NGP, show that incorporating depth supervision with our planar architectural depth priors improves visual quality compared to methods that rely solely on photometric loss. Moreover, our proposed depth loss for boundary surfaces outperforms Mean Squared Error (MSE) loss on both datasets, yielding superior visual coherence. Additionally, integrating our patch-based depth regularization techniques further refines results, enhancing depth consistency across the scene. Last but not least, training with depth supervision using our depth priors accelerates the process, further enhancing the efficiency of our approach.

In summary, the main contributions of this work are as follows:

- The generation of a synthetic 360-degree indoor dataset, comprising two distinct scenes, which we intend to make publicly available to support future research.
- The design of an algorithm for producing dense depth priors on planar architectural surfaces, such as walls, ceilings, and floors.
- The formulation of a new depth loss function tailored for these planar boundary surfaces.
- The development of a patch-based depth regularization technique, incorporating bilateral and joint bilateral filters.

## 2 RELATED WORK

Research to enhance NeRFs rendering quality has led to various *depth regularization* and *depth supervision* methods aimed at improving rendering quality by refining the scene’s geometry.

*Implicit regularization* approaches leverage pre-trained models to encode geometry and appearance priors. For instance, Pixel-NeRF (Yu et al., 2021) directly integrates features from a convolutional neural network (CNN) trained on multiple scenes to condition the NeRF model, while DietNeRF (Jain et al., 2021) incorporates a regularization term in its loss function to enforce consistency between high-level features across both known and novel views. However, these regularization methods often struggle when applied directly to indoor datasets due to domain gaps, as the CNNs are typically pre-trained on ImageNet (Deng et al., 2009), which predominantly features natural images. Bridging this gap can

be resource-intensive and may require additional fine-tuning (Chen et al., 2023).

*Explicit regularization* methods specifically target high-frequency artifacts by smoothing inconsistencies between adjacent regions. RegNeRF (Niemeyer et al., 2022), for example, enforces similarity constraints on neighboring pixel patches, while InfoNeRF (Kim et al., 2022) minimizes a ray entropy model to maintain consistent ray densities across views.

Although regularization techniques can enhance rendering quality to some degree, their overall impact remains limited (Chen et al., 2023). In contrast, *depth supervision* addresses sparse scenarios and regions with less prominent visual features by providing a stronger optimization signal through an additional depth constraint that leverages depth priors and ensures consistency between rendered and ground truth depth (Rabby and Zhang, 2023). For instance, DS-NeRF (Deng et al., 2022) and Urban-NeRF (Rematas et al., 2022) incorporate a depth loss that adjusts the predicted depth to match available sparse depth data.

In the context of indoor scene synthesis, notable research efforts such as Dense Depth Priors (Roessle et al., 2022) and NerfingMVS (Wei et al., 2021) have proposed methods to enhance NeRF performance by transforming sparse data points—typically a byproduct of the Structure from Motion preprocessing step used for estimating camera poses—into dense depth maps using a monocular depth completion model. In the first approach, these dense depth priors are leveraged to guide the NeRF optimization process, effectively accounting for uncertainty in depth estimation while minimizing the error between predicted and true depth values (Roessle et al., 2022). NerfingMVS (Wei et al., 2021) builds on this by calculating loss through comparisons between rendered depth and learned depth priors, incorporating confidence maps to weigh the reliability of the depth estimates. These supervision strategies generally yield superior results compared to those relying solely on sparse depth points (Wang et al., 2023a). However, their limitation lies in a lack of view consistency, as each view is processed individually during the depth completion step. StructNeRF (Chen et al., 2023) addresses this by incorporating photometric consistency, comparing source images with their warped counterparts from other viewpoints in visually rich regions. To handle non-textured areas, it introduces a regularization loss that enforces planar consistency, encouraging points within regions identified by planar segmentation masks to lie on a single plane. This approach helps maintain multi-view consistency, though the warping process significantly increases computational cost (Wang et al., 2023c). Notably, methods

that utilize depth supervision struggle in areas with low visual features, either because they inherit limitations from Structure from Motion or Multi-View Stereo depth estimates, or, as in the case of StructNeRF, rely on warping for photometric consistency.

Research on 360-degree panorama NeRF-based view synthesis, similar to the pinhole camera model, widely applies additional depth supervision for optimization (Gu et al., 2022), (Wang et al., 2023b), (Kulkarni et al., 2023). While PERF (Wang et al., 2023b) estimates depth using a 360-degree depth estimator, Omni-NeRF (Gu et al., 2022) and 360Fusion-NeRF (Kulkarni et al., 2023) derive depth maps by projecting 2D image pixels onto a spherical surface and analyzing the intersections of rays with the scene geometry from multiple views. However, since our work involves images prior to their assembly into a 360-degree panorama and adheres to the pinhole camera model, research focused on spherical projections is not directly related to our scenario.

Compared to previous methods, our approach to computing architectural priors for indoor scenes and utilizing boundary loss shares similarities with Dense Depth Priors (Roessle et al., 2022) and NerfingMVS (Wei et al., 2021), as it follows the depth supervision approach using depth maps. Similar to StructNeRF (Chen et al., 2023), we employ fundamental architectural principles to address non-textured areas. Unlike other studies, our approach imposes reliable geometric constraints in featureless regions of large architectural planes, without dependence on the inaccuracies associated with photometric consistency in these areas, thereby efficiently and effectively tackling challenges in low-feature regions. Our depth regularization technique shares conceptual similarities with RegNeRF (Niemeyer et al., 2022) in its use of patches. However, our approach not only smooths out noise but also better preserves edges, enhancing depth consistency without sacrificing structural detail.

### 3 DATASET

Our focus is on capturing indoor scenarios using a 360-degree camera. To cover the entire space, we recommend a mounted, movable setup. Rather than working with a stitched 360-degree panorama, we propose using a series of individual raw views prior to their assembly (see Figure 1). While unconventional, this approach has the potential to significantly improve the accuracy of extrinsic and intrinsic data compared to a stitched panorama—essential for the NeRF pipeline—and, consequently, enhance the overall quality of NeRF-rendered scenes (Gu et al., 2022).

Hence, we generated a custom dataset in Blender comprising two synthetic indoor scenes: a bedroom (6×8×3.8 m) and a living room (10×10×3.4 m). Both scenes are modeled with the floor at  $Z = 0$  and the origin at the center of the floor, with orthogonal coordinate axes and the positive  $Z$ -axis extending upward. Individual images of an unstitched 360-degree panorama were captured using Blender’s perspective camera with a 27° horizontal and 40° vertical field of view. Each 360-degree horizontal sweep consisted of 15 images, spaced at 24° intervals, with 5 additional images covering the ceiling by first rotating the camera upward and tilting it in four directions. A 3° overlap between adjacent images ensured seamless assembly. Cameras were positioned in a grid pattern across the scene with random noise added for realism.

The living room dataset comprises 1200 training images and 540 evaluation images, while the bedroom dataset includes 840 training images and 300 evaluation images. Each RGB image, at a resolution of 1080×1920 pixels, is provided with camera parameters, depth maps, and segmentation maps for planar architectural surfaces such as floors, ceilings, and walls.

## 4 METHODOLOGY

In this section, we outline our methodology for enhancing NeRF rendering quality in indoor environments, specifically focusing on reducing cloudy artifacts, commonly called “floaters”, that often appear on featureless surfaces. Our approach incorporates custom depth estimation techniques for planar architectural surfaces, such as walls, floors, and ceilings, along with a loss function tailored for boundary regions. Moreover, we propose a depth regularization technique that complements the previous approach by refining rendering quality across the entire scene.

We begin by discussing depth supervision techniques, followed by an introduction of a novel depth estimation method explicitly designed for planar architectural surfaces in indoor scenes. Next, we introduce a boundary loss function that enforces spatial constraints, improving depth accuracy along architectural boundaries. Finally, we outline our custom patch-based depth regularization method.

### 4.1 Depth Supervision

Depth supervision is an effective approach to mitigate floating artifacts by comparing rendered and ground truth depth (Wang et al., 2023a). It constrains the density distribution, enforcing geometric consistency.

Specifically, the color  $\hat{C}(\mathbf{r})$  and depth  $\hat{D}(\mathbf{r})$  of a pixel along a ray  $\mathbf{r}$  are rendered by NeRFs as follows:

$$\hat{C}(\mathbf{r}) = \sum_{i=1}^N w_i \mathbf{c}_i, \quad (1)$$

$$\hat{D}(\mathbf{r}) = \sum_{i=1}^N w_i t_i, \quad (2)$$

where  $\hat{C}(\mathbf{r})$  is the final color rendered for the pixel along ray  $\mathbf{r}$ , and  $\hat{D}(\mathbf{r})$  is the estimated depth from the camera to the pixel along ray  $\mathbf{r}$ . Here,  $N$  denotes the number of samples along  $\mathbf{r}$ .

The weight for the  $i$ -th sample, representing the contribution of a sample  $i$  along the ray  $\mathbf{r}$  to the final color and depth values for the corresponding pixel, is defined as:

$$w_i = T_i \alpha_i. \quad (3)$$

The transmittance  $T_i$  at sample  $i$ , indicating the probability of light reaching the sample unimpeded, is defined as:

$$T_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j \Delta_j\right). \quad (4)$$

The opacity  $\alpha_i$  at sample  $i$  represents the likelihood that light is absorbed or scattered at sample  $i$  and is given by:

$$\alpha_i = 1 - \exp(-\sigma_i \Delta_i). \quad (5)$$

Further,  $\sigma_i$  is the volume density at sample  $i$  and  $\Delta_i = t_{i+1} - t_i$  is the distance between adjacent samples. Here,  $\mathbf{c}_i$  represents the RGB color, and  $t_i$  is the distance from the camera origin to the  $i$ -th sample.

NeRFs are optimized by enforcing rendered color consistency through a photometric loss function, commonly defined as the Mean Squared Error (MSE) between the rendered and ground truth pixel colors (Rabby and Zhang, 2023):

$$\mathcal{L}_{\text{color}} = \sum_{\mathbf{r} \in \mathcal{R}} |\hat{C}(\mathbf{r}) - C(\mathbf{r})|_2^2, \quad (6)$$

where  $\mathcal{R}$  represents the set of rays in each training batch, and  $C(\mathbf{r})$  and  $\hat{C}(\mathbf{r})$  denote the ground truth and predicted RGB colors for each ray  $\mathbf{r}$ , respectively.

Depth supervision is applied by combining this photometric loss with an additional depth loss:

$$\mathcal{L} = \lambda_{\text{color}} \mathcal{L}_{\text{color}} + \lambda_{\text{depth}} \mathcal{L}_{\text{depth}}, \quad (7)$$

where  $\lambda_{\text{color}}$  and  $\lambda_{\text{depth}}$  are weighting factors that balance the contributions of the photometric and depth losses, respectively.

In this work, we utilize an MSE loss to compare the rendered and ground truth depths:

$$\mathcal{L}_{\text{depth}} = \sum_{\mathbf{r} \in \mathcal{R}} |\hat{D}(\mathbf{r}) - D(\mathbf{r})|_2^2. \quad (8)$$



Figure 1: Raw images captured with a pinhole camera model, showing unstitched frames prior to assembly into a 360-degree panorama. The living room is depicted in the first two rows and the bedroom in the last two. The first 15 images (from top left to bottom right) depict a 360-degree horizontal sweep, while the final 5 images capture the upper surroundings.

Here,  $D(\mathbf{r})$  and  $\hat{D}(\mathbf{r})$  are the ground truth and predicted depths, respectively, for ray  $\mathbf{r}$  from the ray batch  $\mathcal{R}$ .

However, depth supervision relies on accurate ground truth depth data, which is often difficult to obtain in real-world scenarios (Ming et al., 2021). A common approach for acquiring depth priors for NeRF is through Structure from Motion techniques, particularly COLMAP, which generates depth information as a byproduct of camera pose estimation (Roessle et al., 2022). Since Structure from Motion methods rely on keypoint matching across multiple images to establish correspondences, they often struggle on textureless areas lacking distinctive visual features—a challenge especially pronounced in indoor environments dominated by uniform architectural surfaces.

## 4.2 Depth Estimation for Planar Architectural Surfaces

We propose a fast, simple, and computationally efficient method to estimate depth in featureless indoor regions such as walls, floors, and ceilings. The approach assumes the Z-axis origin is calibrated to lie on the floor plane. If not, three non-collinear camera positions at a constant height (e.g., tripod-mounted) must be available. Room height must also be known, along with semantic segmentation for wall, floor, and ceiling classes, which can be efficiently generated using pretrained models such as DeepLab (Chen et al., 2018).

Depth computation leverages the NeRF ray representation (Mildenhall et al., 2021), defined as:

$$\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}_{\text{unit}}, \quad (9)$$

where  $\mathbf{o}$  is the camera origin and  $\mathbf{d}_{\text{unit}}$  is a unit vector representing the ray direction. Using camera parameters, the Euclidean depth  $t$  of a pixel  $\mathbf{P}$  is determined by setting one known component of its 3D world coordinate (e.g., the Z-coordinate  $P_z$ , which represents

height in 3D space). Knowing  $t$  allows the recovery of the remaining 3D coordinates of  $\mathbf{P}$ .

For floors and ceilings, if the Z-axis origin lies on the floor, floor depth can be computed directly by setting  $P_z = 0$ , while ceiling depth is computed by setting  $P_z$  to the ceiling height. Without calibration, the plane equation  $s_{\text{cam}}$  is derived from three non-collinear camera positions. Parallel planes are then calculated at distances equal to the camera height above and below  $s_{\text{cam}}$ . The floor plane  $s_{\text{floor}}$  is identified as the parallel plane that intersects the ray corresponding to an arbitrary floor pixel. Next, the ceiling plane  $s_{\text{ceil}}$  is determined similarly, accounting for the ceiling height relative to the camera.

To estimate wall depths, border pixels where walls meet the ceiling and floor are first identified. Three non-collinear points (two from one border and one from the other) are selected to define the wall plane  $s_{\text{wall}}$ . Finally, depth for walls, ceiling, and floor is computed as the Euclidean distance from the ray origins of pixels belonging to the corresponding segmentation classes to their intersection points with the respective planes.

### 4.3 Boundary Loss for Architectural Surfaces

When a ray travels through open space within a room and does not intersect any surface, its transmittance  $T_i$  remains high, meaning the ray continues unimpeded through the scene, while its opacity  $\alpha_i$  remains low, reflecting the absence of intersecting material. This combination of high transmittance and low opacity produces low weights along the ray’s path (see eq. (3)), as there is minimal interaction to indicate boundaries, as depicted by the yellow downward arrows in Figure 2.

However, as the ray reaches a boundary surface (like a wall or ceiling), the interaction characteristics change. The transmittance  $T_i$  remains high initially, as the ray is still progressing through space, but the opacity  $\alpha_i$  begins to rise due to the increasing material density encountered at the boundary. As illustrated by the upward blue arrow in Figure 2, this increase in opacity correlates with higher weights near the boundary, highlighting the role of these architectural surfaces in defining the spatial limits within the scene. When the ray finally intersects a boundary surface, the weights along the ray peak, often reaching a maximum (e.g., a weight of 1), as the ray’s traversal is effectively complete (Szeliski, 2022).

Based on these observations, we introduce a boundary loss function that leverages our architec-

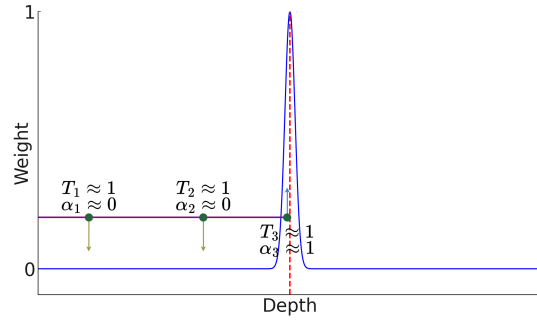


Figure 2: Illustration of a Gaussian distribution modeling the weight  $w_i$  along a ray which hits the boundary surface (e.g., a wall) depicted by the red dotted line. The purple solid line indicates the ray with the green dots representing samples.

tural depth priors:

$$\mathcal{L}_{\text{bound}} = \sum_{\mathbf{r} \in \mathcal{R}} \sum_t \left( w_i - e^{-\left( \frac{(t_i - D(\mathbf{r}))^2}{2\sigma^2} \right)} \right)^2, \quad (10)$$

where  $D(\mathbf{r})$  is the ground truth depth of ray  $\mathbf{r}$  from ray batch  $\mathcal{R}$ ,  $w_i$  is a weight corresponding to a point, which is sampled on the ray  $\mathbf{r}$  at the distance  $t_i$  from the ray origin.

For pixels corresponding to architectural surfaces, the boundary loss penalizes weights of samples far from the surface and boosts weights close to it, enforcing correct architectural constraints.

### 4.4 Patch-Based Depth Regularization

To complement our depth supervision on planar architectural surfaces and mitigate rendering irregularities beyond these regions, we draw inspiration from Reg-NeRF (Niemeyer et al., 2022) and propose a depth regularization method that operates on image patches. This approach promotes smooth and consistent depth predictions across rendered views, effectively reducing noise and artifacts while preserving essential structural details. Specifically, we apply a bilateral (Tomasi and Manduchi, 1998) or joint bilateral filter (He et al., 2012) to regularize the depth within each patch.

**Filtering the Depth Patch.** We begin by applying a bilateral or joint bilateral filter to a rendered depth patch  $\hat{D}(p)$ , where  $p$  is a patch from the set  $\mathbf{P}$ . The bilateral filter accounts for both spatial proximity and depth similarity, while the joint bilateral filter additionally considers intensity similarity in the corresponding RGB image. This method ensures that the smoothing of depth values respects the structural edges present in the image.

**Computing the Regularization Loss.** For each depth patch  $p \in \mathbf{P}$ , we calculate the Mean Squared Error (MSE) between the original rendered depth patch  $\hat{D}(p)$  and the filtered depth patch  $\mathcal{F}(\hat{D}(p))$ . We then compute the average of these MSE losses across all patches in  $P$  to obtain a single regularization term:

$$\mathcal{L}_{\text{reg}} = \frac{1}{|\mathbf{P}|} \sum_{p \in \mathbf{P}} \frac{1}{|p|} \sum_{i,j} (\hat{D}(p_{ij}) - \mathcal{F}(\hat{D}(p))_{ij})^2. \quad (11)$$

This regularization term is incorporated into the total loss function in the same manner as depth supervision (see eq. 7).

## 5 EXPERIMENTS

To evaluate the effectiveness of our approach in enhancing NeRF rendering quality in indoor environments, we conducted a series of experiments using Instant-NGP (Müller et al., 2022) within the Nerfstudio framework (Tancik et al., 2023). Instant-NGP was chosen for its hash encoding, which captures objects of varying sizes, and occupancy grids, which focus computation on meaningful areas in indoor scenes with significant empty space. Given computational constraints, we downsampled our datasets by a factor of two, resulting in a final resolution of 540x960 pixels. All models were trained on an NVIDIA GeForce RTX 3090 GPU using the Adam optimizer with parameters  $\beta_1 = 0.9$ ,  $\beta_2 = 0.99$ , and  $\epsilon = 10^{-8}$ . Neither weight decay nor gradient clipping was applied. We set the hash table size  $T$  to 22, the maximum resolution  $N_{\text{max}}$  to 32,768, the density MLP to a depth of 2 and a width of 64, and the color MLP to a depth of 2 and a width of 128. These values were determined through hyperparameter optimization. All other hyperparameters followed Nerfstudio defaults. It took 200,000 iterations to train the models with photometric loss on both scenes, with performance plateauing beyond this point.

**Patch-Based Depth Regularization.** We implemented patch-based regularization using the open-source library Kornia, utilizing its default parameters for bilateral and joint bilateral filters: a  $9 \times 9$  kernel size, range sigma ( $\sigma_{\text{color}}$ ) of 10 to control intensity similarity, and spatial sigma ( $\sigma_{\text{space}}$ ) of  $75 \times 75$  to define the spatial extent of the filter.

Training with patches significantly extended the process, requiring additional time for the network to capture global image structure. To address this, the model was first trained to convergence with photometric loss, followed by patch-based regularization to refine details. We trained with a patch size of 16, as larger patches (32 and 64) remained undertrained

even after 400,000 iterations and significantly increased training time. The best results were achieved with  $\lambda_{\text{color}} = 1$  and  $\lambda_{\text{reg}} = 10^{-7}$  for both bilateral and joint bilateral loss. Models using these parameters converged in 280,000 iterations. For comparison, we implemented patch similarity constraints as described in RegNeRF (Niemeyer et al., 2022), following the same training strategy.

**Depth Supervision with Planar Architectural Depth Priors.** As a preprocessing step, we computed depth estimates for the floor, ceiling, and walls using semantic segmentation generated in Blender. To evaluate the accuracy of these depth priors, we compared them against Blender-generated depth maps as ground truth. The results demonstrated high accuracy, with Root Mean Square Error (RMSE) values of 2.786 mm for the bedroom scene and 3.201 mm for the living room scene.

Next, we incorporated these architectural depth priors into the training process. Instant-NGP was trained on both scenes with depth supervision, employing MSE and BoundL loss alongside the priors, continuing each model until convergence. For models utilizing BoundL, we set  $\delta$  to 1 mm. The weight values for the losses were set to  $\lambda_{\text{color}} = 10$  and  $\lambda_{\text{depth}} = 10$  for the bedroom scene, and  $\lambda_{\text{color}} = 1$  and  $\lambda_{\text{depth}} = 1$  for the living room scene. Depth-supervised models converged in only 120,000 iterations, demonstrating the efficiency of incorporating planar architectural depth priors into the training process.

## 6 RESULTS

Renderings produced by our baseline model, which relies solely on photometric loss, confirm our initial observation: "floaters" are more common on textureless surfaces like walls, floors, and ceilings (see Figure 4). In contrast, objects with rich visual features—such as plants, books, and paintings—exhibit fewer floaters, as shown in Figure 3. Notably, cloudy artifacts consistently align with incorrect depth estimations. This outcome underscores the limitations of NeRFs when relying solely on RGB optimization signals to accurately predict geometric constraints in featureless regions. Interestingly, some inconsistencies in the rendered depth maps did not produce visible artifacts in the color image, indicating a degree of tolerance in NeRF’s volume rendering.

Visual observations reveal a noticeable reduction in artifacts for depth-guided methods compared to those without depth supervision (see Figure 5). Moreover, the BoundL loss demonstrates fewer artifacts

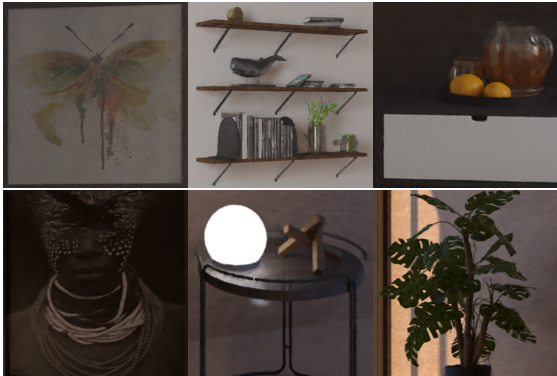


Figure 3: Renderings with Instant-NGP trained on our 360-degree indoor dataset using photometric loss show high visual fidelity on detail-rich areas.

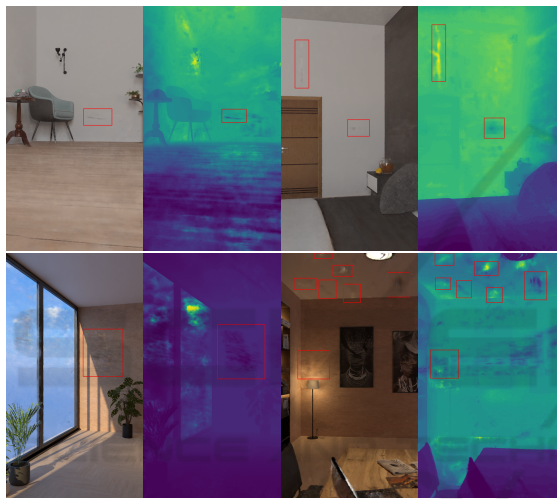


Figure 4: Renderings produced by Instant-NGP trained on our 360-degree indoor dataset with photometric loss are displayed alongside their corresponding depth maps. Red bounding boxes highlight floaters in front of walls, ceilings, or floors, caused by incorrect depth estimations.

than MSE loss, producing cleaner and more accurate renderings (see Figure 6). This is likely due to BoundL’s ability to directly address the weights of samples, effectively reducing ambiguity during the volume rendering procedure.

To quantitatively compare our models, we employ standard view synthesis evaluation metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS). As expected, models with depth supervision outperform their counterparts, with BoundL loss (both with and without joint-bilateral regularization) achieving the highest metrics.

Our patch-based regularization methods deliver consistent quality improvements across both indoor scenes (see Table 1), achieving better metrics com-



(a) RGB (b) BoundL (c) RGB (d) BoundL



(e) RGB (f) BoundL (g) RGB (h) BoundL

Figure 5: Renderings with Instant-NGP trained on our 360-degree indoor dataset, using photometric loss in (a), (c), (e), and (g), and depth supervision with BoundL in (b), (d), (f), and (h). Red bounding boxes highlight floaters, which are minimized through depth guidance with planar architectural depth priors and BoundL.

pared to RegNeRF’s depth patch regularization. This advantage is likely due to the ability of bilateral and joint-bilateral filtering to reduce noise while preserving sharp edge transitions and essential structural details. Moreover, joint-bilateral regularization demonstrates additional gains over the bilateral approach.

Notably, performance metrics vary across scenes, with the living room consistently outperforming the bedroom. This is likely due to obstructions in the bedroom—such as the large bed—limiting ray coverage in occluded areas.

Further, depth-supervised models also demonstrate faster convergence, requiring only 120,000 iterations compared to 200,000 for models trained solely on RGB loss (see Figure 6), and 280,000 iterations for those using patch-based depth supervision. This speedup is attributed to depth supervision, which enables the model to quickly identify empty spaces, concentrate sampling on occupied regions, and provide a stronger optimization signal (Deng et al., 2022).

## 7 CONCLUSIONS

This research tackles the challenge of textureless regions for NeRF-based novel view synthesis in indoor environments. To address this, we developed a depth guidance approach for large planar surfaces, such as walls, floors, and ceilings—regions where



Table 1: Quantitative comparison for 360-degree indoor scenes on the evaluation dataset. We report PSNR, SSIM and LPIPS. "Arch. planar" refers to depth-guided methods that utilize depth priors for architectural planar surfaces.

Method	Bedroom Scene			Livingroom Scene		
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
Only RGB loss	31.163	0.749	0.378	34.008	0.856	0.268
Arch. planar + MSE	34.174	0.762	0.327	36.790	0.869	0.252
Arch. planar + BoundL	34.309	<b>0.792</b>	0.285	36.902	<b>0.921</b>	0.248
RegNeRF patch	30.830	0.750	0.361	34.033	0.858	0.259
Bilateral filter	31.890	0.750	0.359	35.200	0.858	0.256
Joint bilateral filter	32.709	0.763	0.342	36.127	0.871	0.256
Arch. planar + BoundL with joint bilateral	<b>34.410</b>	0.765	<b>0.281</b>	<b>36.935</b>	0.898	<b>0.245</b>

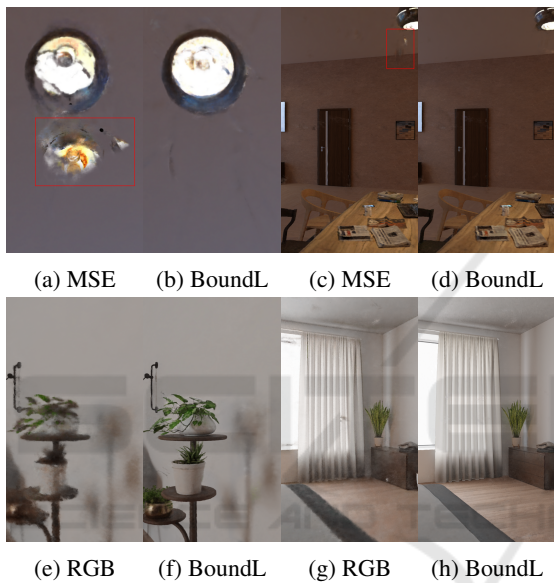


Figure 6: Renderings with Instant-NGP trained on our 360-degree indoor dataset: (a)-(d) compare depth supervision methods, with MSE in (a) and (c), and BoundL in (b) and (d). Red boxes highlight areas where MSE results exhibit rendering artifacts that BoundL successfully mitigates. (e)-(h) illustrate the faster convergence of models with depth supervision, showing BoundL examples in (f) and (h) and photometric loss in (e) and (g).

NeRFs often struggle. Specifically, we proposed an efficient method to compute depth priors for the mentioned surfaces and introduced a depth loss function, BoundL, to enforce depth constraints on these planar boundaries. This is complemented by our patch-based regularization, which utilizes bilateral and joint-bilateral filtering.

To evaluate our approach, we created a synthetic indoor dataset comprising two distinct scenes that simulate individual views within a 360-degree panorama prior to assembly. Working with a series of raw images captured with a pinhole camera model aids in determining accurate image poses, eliminating

the need to account for geometric distortions in the final 360-degree stitched panorama.

Our results demonstrate clear improvements in rendering quality, both visually and quantitatively, when incorporating our planar depth priors with depth supervision through MSE and BoundL loss. Notably, BoundL consistently outperforms MSE across both scenes. Additionally, our patch regularization techniques surpass RegNeRF’s patch depth constraints, yielding subtle yet stable quantitative gains.

With all enhancements enabled, we achieved an increase in PSNR of up to 3 dB compared to the baseline model using only photometric loss. These improvements underscore the robustness and effectiveness of our approach in refining NeRF rendering for complex indoor environments.

Future work will extend our methods to real-world data, with optimizations to account for noisy camera parameters. Additionally, incorporating sparse depth data from feature-rich regions and enforcing strict planarity on other linear surfaces could further improve model accuracy and rendering quality.

## ACKNOWLEDGEMENTS

This work has partly been funded by the German Federal Ministry for Digital and Transport (project EConoM, grant no 19OI22009C) and the German Federal Ministry of Education and Research (project VoluProf, grant no. 16SV8705).

## REFERENCES

- Arandjelović, R. and Zisserman, A. (2021). Nerf in detail: Learning to sample for view synthesis. *arXiv:2106.05264*.
- Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. on Pat-*

- tern Analysis and Machine Intelligence*, 39(12):2481–2495.
- Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. *arXiv:1706.05587*.
- Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*.
- Chen, Z., Wang, C., Guo, Y.-C., and Zhang, S.-H. (2023). Structnerf: Neural radiance fields for indoor scenes with structural hints. *IEEE Trans. on Pattern Analysis and Machine Intelligence*.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 248–255.
- Deng, K., Liu, A., Zhu, J.-Y., and Ramanan, D. (2022). Depth-supervised nerf: Fewer views and faster training for free. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12882–12891.
- Gasperini, S., Morbitzer, N., Jung, H., Navab, N., and Tombari, F. (2023). Robust monocular depth estimation under challenging conditions. In *Proc. IEEE/CVF Int. Conf. on Computer Vision*, pages 8177–8186.
- Gu, K., Maugey, T., Knorr, S., and Guillemot, C. (2022). Omni-nerf: neural radiance field from 360 image captures. In *IEEE Int. Conf. on Multimedia and Expo (ICME)*.
- He, K., Sun, J., and Tang, X. (2012). Guided image filtering. *IEEE transactions on pattern analysis and machine intelligence*, 35(6):1397–1409.
- Jain, A., Tancik, M., and Abbeel, P. (2021). Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proc. IEEE/CVF International Conference on Computer Vision*, pages 5885–5894.
- Kim, M., Seo, S., and Han, B. (2022). Infonerf: Ray entropy minimization for few-shot neural volume rendering. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12912–12921.
- Kulkarni, S., Yin, P., and Scherer, S. (2023). 360fusionnerf: Panoramic neural radiance fields with joint guidance. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7202–7209.
- Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. (2021). Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106.
- Ming, Y., Meng, X., Fan, C., and Yu, H. (2021). Deep learning for monocular depth estimation: A review. *Neurocomputing*, 438:14–33.
- Müller, T., Evans, A., Schied, C., and Keller, A. (2022). Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. on Graphics (TOG)*, 41(4):1–15.
- Niemeyer, M., Barron, J. T., Mildenhall, B., Sajjadi, M. S., Geiger, A., and Radwan, N. (2022). Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5480–5490.
- Rabby, A. and Zhang, C. (2023). Beyondpixels: A comprehensive review of the evolution of neural radiance fields. *arXiv:2306.03000*.
- Rematas, K., Liu, A., Srinivasan, P. P., Barron, J. T., Tagliasacchi, A., Funkhouser, T., and Ferrari, V. (2022). Urban radiance fields. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12932–12942.
- Roessle, B., Barron, J. T., Mildenhall, B., Srinivasan, P. P., and Nießner, M. (2022). Dense depth priors for neural radiance fields from sparse input views. In *Proc. Int. Conf. on Computer Vision and Pattern Recognition*.
- Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Proc. 18th Int. Conf. on Medical image computing and computer-assisted intervention (MICCAI)*, pages 234–241.
- Szeliski, R. (2022). *Computer vision: algorithms and applications*. Springer Nature.
- Tancik, M., Weber, E., Ng, E., Li, R., Yi, B., Wang, T., Kristoffersen, A., Austin, J., Salahi, K., Ahuja, A., et al. (2023). Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–12.
- Tomasi, C. and Manduchi, R. (1998). Bilateral filtering for gray and color images. In *Proc. IEEE/CVF International Conference on Computer Vision*.
- Wang, C., Sun, J., Liu, L., Wu, C., Shen, Z., Wu, D., Dai, Y., and Zhang, L. (2023a). Digging into depth priors for outdoor neural radiance fields. In *Proc. 31st ACM Int. Conference on Multimedia*, pages 1221–1230.
- Wang, G., Wang, P., Chen, Z., Wang, W., Loy, C. C., and Liu, Z. (2023b). Perf: Panoramic neural radiance field from a single panorama. *arXiv:2310.16831*.
- Wang, J., Wang, P., Long, X., Theobalt, C., Komura, T., Liu, L., and Wang, W. (2022). Neuris: Neural reconstruction of indoor scenes using normal priors. In *European Conference on Computer Vision*, pages 139–155. Springer.
- Wang, Y., Xu, J., Zeng, Y., and Gong, Y. (2023c). Anisotropic neural representation learning for high-quality neural rendering. *arXiv:2311.18311*.
- Wei, Y., Liu, S., Rao, Y., Zhao, W., Lu, J., and Zhou, J. (2021). Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In *Proc. IEEE/CVF International Conference on Computer Vision*, pages 5610–5619.
- Yu, A., Ye, V., Tancik, M., and Kanazawa, A. (2021). pixelnerf: Neural radiance fields from one or few images. In *Proc. IEEE/CVF conference on computer vision and pattern recognition*, pages 4578–4587.