RoomNeRF: Representing Empty Room as Neural Radiance Fields for View Synthesis

ManKyu Kong
mangyu0929@yonsei.ac.kr
Seongwon Lee
won4113@yonsei.ac.kr
Euntai Kim
etkim@yonsei.ac.kr

School of Electrical and Electronic Engineering
Yonsei University
Seoul, Korea

Abstract

We present a method for novel view synthesis of empty rooms from object-existing room images. Despite the remarkable achievements of previous inpainted NeRFs for object removal tasks, they have a limitation in completely reconstructing the empty room due to the lack of consideration for the room's characteristics. Our proposed network, named RoomNeRF, is designed to fully exploit the shared intrinsic properties of each plane of the room via the Pattern Transfer (PT) and Planar Constraint (PC). For each plane, the PT and PC modules capture shared visual patterns and geometrical structures, respectively, and transfer them to areas occluded by objects, enabling realistic empty room reconstructions without being disturbed by invisible areas of the input images. With these internal learning strategies, RoomNeRF successfully performs novel view synthesis of an empty room from multi-object presence images in extensive experiments and demonstrates its superiority.

1 Introduction

Novel view synthesis from a sparse set of images is a long-standing task in computer vision, which is essential for many AR/VR applications. This novel view synthesis allows users to render scenes from static images, virtually enter the environment, and move freely within 3D space. A representative space where the novel view synthesis can be applied is the room. In this case, users can freely look around, explore, and place furniture in a 3D reconstructed empty room in a virtual environment. To accomplish this, we need a novel view synthesis technique along with removing objects present in the room.

Recently, Neural Radiance Fields (NeRF) [19] has achieved astonishing results in novel view synthesis. Following its initial development, various studies have been conducted to explore extensive applications [2, 3, 7, 17, 21, 29, 32, 37]. One of the representative applications is inpainted NeRF [15, 20, 31], which performs novel view synthesis while removing certain objects in the input images. At first glance, these inpainted NeRF solutions seem to be able to perform novel view synthesis of an empty room successfully. However, despite their impressive performance in object-wise removal tasks, these methods face challenges in

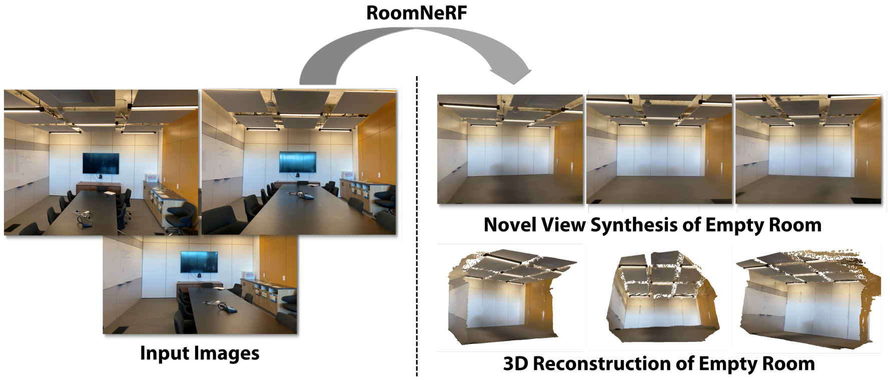
RoomNeRF
Input Images
Novel View Synthesis of Empty Room
3D Reconstruction of Empty Room

Figure 1: Novel view synthesis of an empty room from object-existing room images.

reconstructing an empty room and have a limitation in generating realistic novel view images of the empty room, due to insufficient consideration of the intrinsic properties of the room.

In this paper, we present a network for novel view synthesis of empty rooms from object-existing room images. Our proposed network, named RoomNeRF, is designed to fully exploit the intrinsic properties shared by each plane of the room on top of the inpainted NeRF. To this end, we propose two modules: Pattern Transfer (PT) and Planar Constraint (PC). The PT and PC modules capture visual patterns and geometrical structures shared by each room plane, respectively, and transfer them to the region occluded by objects. This information propagation dramatically improves the reconstruction quality of object-occluded regions in the input images and allows the network to generate realistic novel view images of empty rooms. The proposed RoomNeRF successfully performs the novel view synthesis of empty rooms in various datasets, from virtual to real, demonstrating its superiority. To summarize, we propose RoomNeRF, which presents a new way to synthesize novel views of empty rooms from object-existing room images. The main contributions of our paper are as follows.

- First, we propose Pattern Transfer (PT), which transfers the visual pattern of each plane of a room to the region occluded by objects to perform a consistent visual reconstruction.
- Second, we propose Planar Constraint (PC), which transfers the geometric structure of each plane of a room to the region occluded by objects to perform a consistent geometric reconstruction.

With the aforementioned contributions, our proposed RoomNeRF successfully performs the novel view synthesis of empty rooms in various datasets, from virtual to real, demonstrating its superiority.

2 Related work

Neural Radiance Field (NeRF). The volumetric rendering approach to novel view synthesis has gained prominence in recent times. Particularly, Neural Radiance Field (NeRF) [19], which utilizes neural networks to model a scene's appearance and 3D structure, has emerged as a promising solution in volumetric rendering approaches. NeRF and its extensions also have brought significant advancements beyond traditional novel view synthesis solutions:

better reconstruction performance [2, 3], faster-rendering speed [21], larger fields [29, 37]. Some solutions directly supervise the scene geometry using sparse depth predicted from Structure-from-Motion (SfM) [7, 26], dense depth completed by the neural network [25] or depth from sensors [11, 24], to achieve more accurate scene reconstruction.

As NeRF faithfully reconstructs the scene as it is, various editing methods [15, 17, 20, 31, 32, 34] have also been proposed to manipulate the scene according to specific preferences. In particular, recent discussions have revolved around techniques such as inpainting NeRF[15, 20, 31], which involves removing objects within a scene and restoring the background. Among these inpainting NeRF methods, notable approaches include Spin-NeRF and Object-removal NeRF. Spin-NeRF [20] utilizes perceptual loss rather than photometric loss, and object-removal NeRF [31] weights confidence score for each image and selects view-consistent images during optimization. These solutions show promising performance, however, they are limited to a single object and paint the geometry of scenes neglecting the detailed structure, and sometimes fail to cleanly remove many objects in structured spaces like a room. In this paper, we propose an empty room reconstruction method named Room-NeRF, which preserves the visual and geometric structure of the room while editing, even when removing many objects such as furniture.

Image Inpainting. Early 2D image inpainting approaches complete the missing regions in an image based on patch-based [1, 6] or nearest neighbor methods [8]. After the advent of deep learning, GAN-based approaches showed remarkable performance with adversarial methods [10, 22, 35, 36] and various architectures [14, 16, 33], resulting in successful outcomes. In our work, we utilize LAMA [28], a method that employs Fourier convolutions and high receptive fields to handle challenging scenarios with large masks. While there exist inpainting approaches that utilize perceptual information such as room layout [12] or assuming the background to be plane [23] for image-based rendering; however, they are not well-suited for achieving realistic 3D reconstruction required for novel view synthesis, which is our primary objective. To enable 3D reconstruction through inpainting methods, we combined LAMA[28] with neural radiance fields[19] to reconstruct empty rooms for novel view synthesis.

3 Preliminaries

In this section, a brief explanation of NeRF and naïve inpainted NeRF will be introduced.

3.1 Neural Radiance Field (NeRF)

With the sparse set of images of a static scene, $\mathrm{I} = \{I_i\}_{i=1}^{n}$, and their corresponding camera intrinsic and extrinsic parameters, NeRF represent a 3D scene with a function $F_\theta : (x,d) \to (c,\sigma)$, which encodes 3D coordinate $x = (x,y,z)$ and viewing direction $d = (\theta,\phi)$ to RGB color c and density σ. Using this function F_θ, the rendered color $\hat{\mathrm{C}}(r)$ projected into each pixel is computed as the integrated color of the sampling points of the pixel-associated ray r:

$$\hat{\mathrm{C}}(r) = \sum_{i=1}^{N} T_i(1-\exp(-\sigma_i\delta_i))c_i, T_i = \exp(-\sum_{j=0}^{i-1}\sigma_j\delta_j), \tag{1}$$

where N is the number of sampling points on each ray, and δ_i is the distance between two adjacent sample points. NeRF aims to minimize the difference between the rendered colors

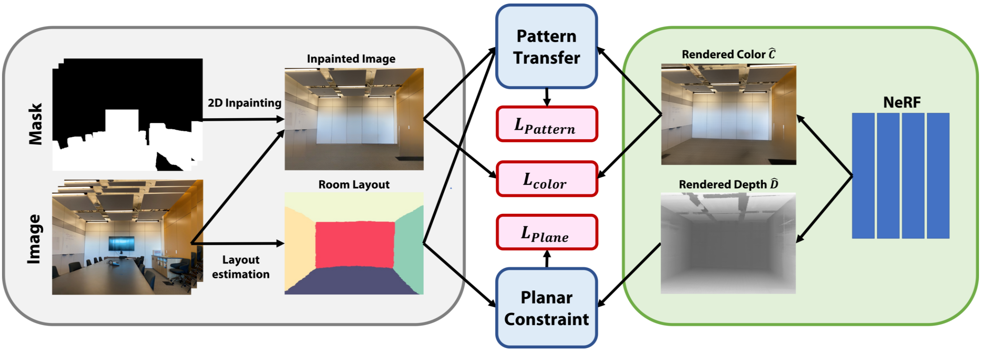

Figure 2: An overview of our proposed method. Our method takes posed RGB images with corresponding masks as input to optimize NeRF representing an empty room. Losses derived from the pattern transfer module and planar constraint module transfer the pattern and the structure of each plane of the room to ensure visual and geometric consistency.

$\hat{\mathrm{C}}(r)$ and the corresponding ground-truth colors $\mathrm{C}(r)$ for all ray $r \in R$:

$$L_{color} = \sum_{r \in R} \|\hat{\mathrm{C}}(r) - \mathrm{C}(r)\|^2, \tag{2}$$

where R is the set of rays used in NeRF training for each step.

Additionally, some NeRF studies [7, 25] have used pseudo sparse depths derived from Structure from Motion (SfM) [26] as a weak supervision. The rendered depth $\hat{D}(r)$ projected corresponding to each pixel r is predicted from the distribution of occupancy on the rays:

$$\hat{\mathrm{D}}(r) = \sum_{i=1}^{N} w_i t_i, w_i = T_i(1 - \exp(-\sigma_i \delta_i)). \tag{3}$$

In this case, NeRF additionally aims to minimize the difference between the rendered depth $\hat{\mathrm{D}}(r)$ and the corresponding ground-truth depth $\mathrm{D}(r)$ for all ray $r \in R$:

$$L_{depth} = \sum_{r \in R} \|\hat{D}(r) - D(r)\|^2. \tag{4}$$

3.2 Naïve Inpainted NeRF

The most intuitive approach to removing objects from NeRF is to train them with object-inpainted images. The inpainting module F_{in} generates inpainted images $\tilde{\mathrm{I}} = \{\tilde{I}_i\}_{i=1}^{n}$ with a set of binary occlusion masks $\mathrm{M} = \{M_i\}_{i=1}^{n}$ obtained through object segmentation, which can be replaced by other methods:

$$\tilde{\mathrm{I}} = F_{in}(I, M). \tag{5}$$

By using these object-inpainted images $\tilde{\mathrm{I}}$ as training data, we can naïvely train a NeRF that implies a scene without objects.

4 Method

Even with high-quality image inpainting methods, inpainting large occluded regions in an image can result in poor inpainting results due to large masks, shadows, or inaccurate masks.

Furthermore, individually inpainted images lack 3D consistency, resulting in blurred and corrupted novel-view images. The same problems also arise when reconstructing empty rooms from images of rooms with many objects. However, we have found that in the case of a room, we can effectively tackle these issues by leveraging planar-internal information.

In this section, we introduce a novel network for novel view synthesis of empty rooms, named RoomNeRF. Our proposed RoomNeRF is designed to fully exploit the shared intrinsic properties of each plane of the room via two novel modules: Pattern Transfer (PT) and Planar Constraint (PC). In Sec. 4.1, Pattern Transfer is introduced first, which captures and transfers the shared visual pattern in each plane to the object-occluded region. In Sec. 4.2, Planar Constraint is additionally introduced, which captures and transfers the 3D geometric structures of each plane to the object-occluded region. With these contributed modules, our RoomNeRF successfully generates novel view images of empty rooms. An overview of our proposed method is shown in Fig. 2.

4.1 Pattern Transfer (PT)

In this subsection, we propose Pattern Transfer (PT), which transfers a shared visual pattern from the visible region to the occluded region. In the case of an empty room, it's reasonable to assume that the same planes within the room share a similar visual pattern. With this assumption, we can infer that the visual pattern in the occluded region closely resembles the pattern present in the unoccluded region of the same plane.

The proposed Pattern Transfer module consists of two steps. The first is a search step in which each patch sampled from the occluded regions of the inpainting image searches similar patches in the unoccluded regions of the original image. The second is a transfer step in which the searched similar visual pattern from the original images is transferred to each occluded region.

Specifically, the search step can be described as follows: for each pixel i in the occluded region M, we extract pixel-surrounding patch representation Q_i. Additionally, for each pixel j in the non-occluded region M^c, we extract pixel-surrounding patch representation K_j. The similarity of two patch representations is measured as cosine similarity:

$$d(Q_i, K_j) = \frac{< Q_i \cdot K_j >}{\|Q_i\| \|K_j\|}. \tag{6}$$

With these representations, we can find the most similar patch representation $K_{j'}$ as follows:

$$j' = \underset{j \in (M^c \cap L)}{\operatorname{argmax}} \; d(Q_i, K_j), \tag{7}$$

where L is the mask of the same plane where pixel i belongs.

The transfer step can be described as follows: by minimizing the difference between the patch representation $\hat{R}_i$ rendered from NeRF and the most similar patch representation $K_{j'}$, for all pixel i in the object-occluded region M:

$$L_{PT} = \|M \odot \{\hat{R}_i - K_{j'}\}\|^2. \tag{8}$$

In practice, since rendering all patches requires a lot of computation, we perform the pattern transfer on a single patch every iteration. And also if the cosine similarity $d(Q_i, K_{j'}) < \theta_{thres}$ of the most-similar patch is under the threshold, then the patch is not transferred, as

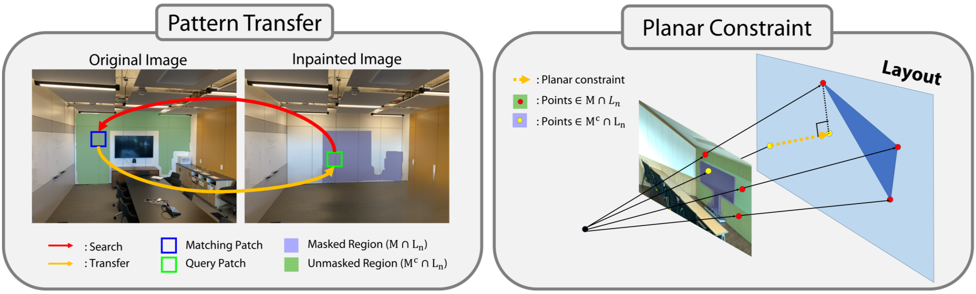

Figure 3: Pattern Transfer (PT) and Planar Constraint (PC).

the most similar patches may not be substantially identical. The detailed illustration of the Pattern Transfer module is illustrated in Fig. 3.

4.2 Planar Constraint (PC)

Simply optimizing NeRF with individually inpainted images not only lacks visual consistency, but it is also geometrically insufficient. Since an empty room is mostly composed of the planar structure as Manhattan world assumption [5], we can assume the occluded area by objects is along the plane of the visible region. With this prior assumption, we can transfer the planar structure from the visible region M^c to the occluded region M.

Specifically, for each plane L_n, we randomly sample three pixels $a_n, b_n, c_n \in M^c \cap L_n$ from visible region M^c. And we unproject the sampled pixels $a_n, b_n, c_n \in p$ to the 3D points $A_n, B_n, C_n \in P$:

$$P = o(r(p)) + \hat{D}(r(p))d(r(p)), \tag{9}$$

where $o(r(p)), d(r(p)), \hat{D}(r(p)) \in \mathbb{R}^3$ is the orientation, direction, and estimated depth with the pixel-associated ray r. With the prior assumption that the occluded region is upon each plane of the visible area, 3D points D_n on the occluded region $M \cap L_n$ also have to be on the plane of visible points $\triangle A_n B_n C_n$.

If 3D point D_n is on the plane $\triangle A_n B_n C_n$ of the visible region, the cross product of $\overrightarrow{A_n D_n}$ and $\overrightarrow{B_n D_n}$ will be perpendicular to $\triangle A_n B_n C_n$ plane. Also, the dot product between the normal vector of the plane $\triangle A_n B_n C_n$ and $\overrightarrow{C_n D_n}$ has to be 0. Thus, minimizing the dot product term forces the 3D points D_n to be on the plane $\triangle A_n B_n C_n$, which works as a loss term:

$$L_{PC} = \frac{1}{N_{Mask}} \sum_{n=1}^{N_{layout}} \sum_{D_n \in M \cap L_n} \left| \overrightarrow{A_n D_n} \times \overrightarrow{B_n D_n} \cdot \overrightarrow{C_n D_n} \right|, \tag{10}$$

where N_{layout} is the number of room layout planes in the image and N_{Mask} is the number of occluded points. The details of Planar Constraint are shown in Fig. 3.

4.3 Objective Function

Our proposed RoomNeRF is trained to minimize the color loss L_{color} and depth loss L_{depth}, which are mentioned at Sec. 3.2. Additionally, we use the pattern transfer loss L_{PT} and planar constraint loss L_{PC} which can help our model reconstruct the 3D empty room scene with consistent visual patterns and geometric structure. Finally, our total loss L_{total} is described as follows:

$$L_{total} = L_{color} + L_{depth} + \lambda_{PT} L_{PT} + \lambda_{PC} L_{PC}. \tag{11}$$

5 Experiments

5.1 Dataset

Existing approaches for novel view synthesis from indoor scenes aim to completely reconstruct the object-existing room with videos or multi-view images. There is, to our knowledge, no standard dataset to evaluate novel view synthesis in an empty room. To address the lack of a dataset, we introduce a real indoor scene dataset. We captured two image sequences with and without objects as training and testing images. Each dataset has about 60 training images and 10 test images captured in the rooms at Yonsei University using iPad 11 Pro. Both our training and test images captured in the identical room share the same intrinsic camera parameter. Their poses are obtained by SfM [26]. For qualitative comparison, we additionally adopted 6 indoor scenes, one scene named 'room' from LLFF dataset [18] and the other five scenes 'room_0', 'room_1', 'room_2', 'office_0' and 'office_3' from replica dataset [27]. We used given ground truth masks of objects on the replica dataset [27].

5.2 Metrics

To evaluate our RoomNeRF, we compare ground-truth images without objects and rendered images from identical camera poses. We use the standard evaluation metrics of original NeRF [19]: Peak Signal-to-Noise Ratio (PSNR) [9], Structural Similarity Index Measure (SSIM) [30], and Learned Perceptual Image Patch Similarity (LPIPS) [38].

5.3 Implementation Details

We built our implementation upon the MPL architecture same as NeRF [19]. We set $\lambda_{PT} = \lambda_{PC} = 0.001$ for the parameters of the loss term. We optimized the model for 50k iterations using Adam optimizer with an initial learning rate of $L_{rate} = 0.0005$. 2048 random rays are sampled for every iteration. For pattern transfer, pre-trained VGG19 is used for extracting feature representation at the activation of $relu3\_1$. Patch size for PT and the similarity threshold θ_{thres} are set as 64x64 and 0.75. The room layouts of each image which are input for PT and PC are estimated through RoomNet [13]. The masks of multiple objects are obtained by using modular interactive video object segmentation [4] with sparse human annotations. The masks are dilated with a 9x9 kernel since the accurate masks of multiple objects allow the edge of objects to cross the boundary letting the shadow and reflections by objects expand.

We compared 'masked NeRF', 'Inpainted NeRF', and object-removing NeRFs [15, 20, 31] on our dataset as a baseline for evaluating novel view synthesis. Masked NeRF optimizes NeRF only with pixels from the unmasked region, ignoring the object region pixels. Inpainted NeRF is trained with images inpainted by LaMa [28]. Since the official codes of object-removal NeRFs are not available, we reproduced them with slight modifications. NeRF-in [15] optimizes with inpainted RGB images and also depth images which are first rendered from pre-trained NeRF and inpainted by LaMa [28]. We reproduced object-removal NeRF [31] which weights uncertainty score for each training image and undergoes a view selection process for every mid-iteration. Since our dataset is absent of data from depth senor, we modified the depth data of object-removal NeRF with the rendered depth as NeRF-in [15].

Table 1: Comparison with baselines for novel view synthesis of an empty room on our dataset

	Seminar room			Office room		
	PSNR(↑)	SSIM(↑)	LPIPS(↓)	PSNR(↑)	SSIM(↑)	LPIPS(↓)
Masked NeRF	21.69	0.8383	0.3781	20.61	0.9387	0.3409
Inpainted NeRF	23.44	0.8672	0.3722	21.07	0.9509	0.3199
NeRF-in [15]	23.35	0.8754	0.3420	21.42	0.8778	0.3310
Object-removal [31]	22.60	0.8591	0.3644	21.26	0.9425	0.3237
SPin-NeRF [20]	22.96	0.8643	0.2454	21.26	0.9467	0.2841
RoomNeRF (Ours)	**23.82**	**0.9148**	**0.1546**	**21.58**	**0.9580**	**0.2434**

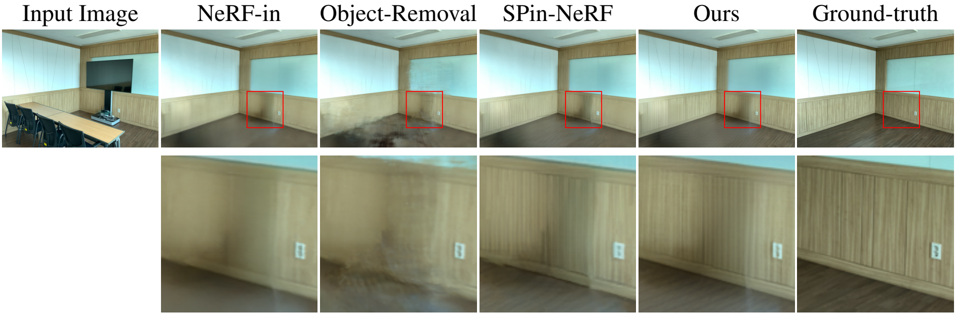

Figure 4: Qualitative comparison with Object-removing NeRFs on our dataset.

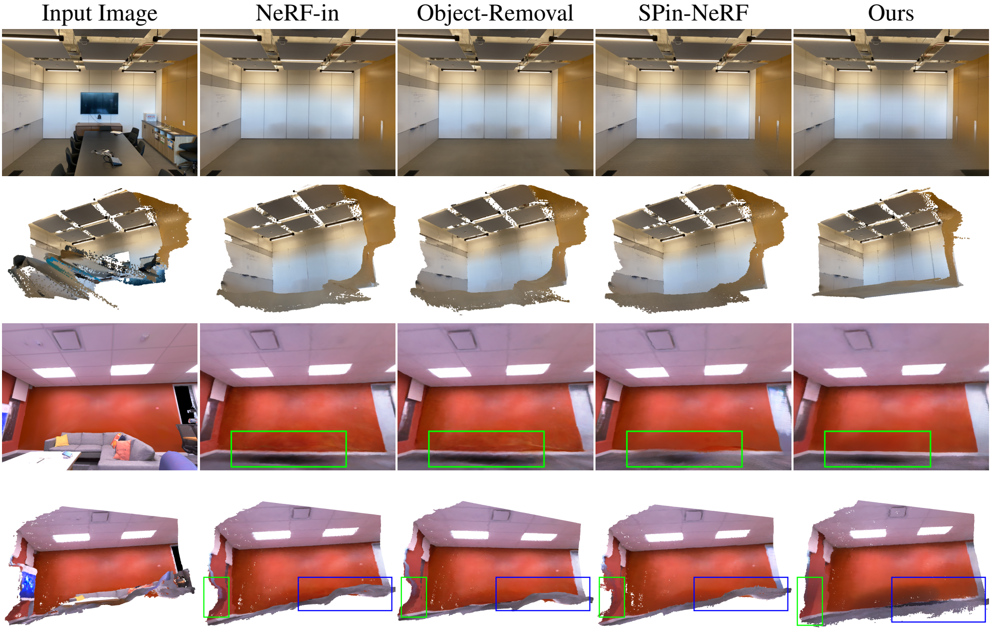

Figure 5: Qualitative results on LLFF Dataset [18] and Replica Dataset [27]

5.4 Experimental Results

Our RoomNeRF is compared against the baselines on our real-world dataset. As shown in the Table. 1, our method is superior to other object-removal NeRFs for novel view synthesis in an empty room. Object-removal NeRFs [15, 20, 31] utilizing depth inpainting are affected

Table 2: Module ablation studies for RoomNeRF.

PT	PC	Seminar room			Office room		
		PSNR(↑)	SSIM(↑)	LPIPS(↓)	PSNR(↑)	SSIM(↑)	LPIPS(↓)
		23.54	0.9048	0.1880	20.65	0.9555	0.2414
✓		23.64	0.9076	0.1673	21.01	0.9574	**0.2254**
	✓	23.70	**0.9155**	0.2250	20.65	0.9565	0.2437
✓	✓	**23.82**	0.9148	**0.1546**	**21.58**	**0.9580**	0.2434

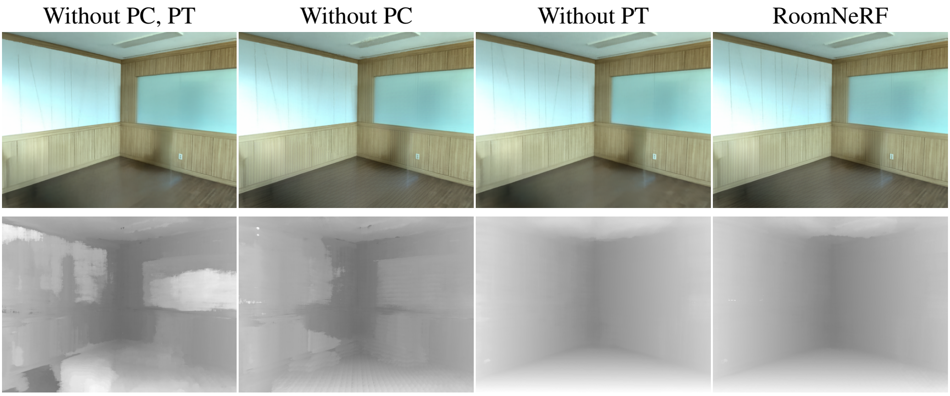

Figure 6: Visual results on ablation studies. The first row shows rendered images from a novel view and the second shows disparity maps.

by the inpainted depth maps generated from the rendered inaccurate depth map, resulting in 3D inconsistency due to different geometry from the real room. On the other hand, our method is similar to the actual wall pattern and structure without being blurred and remaining 3D consistent.

Qualitative Results. In Fig. 4, we show a qualitative comparison with baselines. Our method generates compelling novel view images for the occluded region, while preserving 3D consistency and detailed pattern which is well-blended with the surrounding texture. In contrast, the novel view image from NeRF-in model which simply inpaints NeRF is blurred due to the 3D inconsistency of inpainted images.

Additionally, we conducted experiments on room scenes of the LLFF and replica dataset. Example novel view images and reconstructed 3D point clouds are shown in Fig. 5. More results of the novel view synthesis of an empty room are in our supplementary material.

Ablation Studies. We evaluated our method by applying the modules one by one. The results are shown in Table. 2. As shown in the results, the model applying both Pattern Transfer and Plane Constraint modules shows higher accuracy. The qualitative result of module ablation is shown in Fig. 6. Wall planes of the room are not properly aligned without PC as shown on the rendered disparity map at the novel view. The existence of the Pattern transfer (PT) module makes a difference in the occluded region with the well-generated texture of walls.

6 Conclusion

In this paper, we propose RoomNeRF, which presents a new way to synthesize novel views of empty rooms from object-existing room images. The proposed network leverages the internal patterns and structures of the room to accurately reconstruct occluded regions of a scene well in an efficient way. To this end, we propose two modules. First, we propose a Pattern Transfer module, which captures the internal patterns of the plane of the room and transfers them to the occluded region. Second, we propose a Planar Constraint module, which captures the 3D structures of the plane of the room and transfers them to the occluded region. With these internal learning strategies, our proposed RoomNeRF successfully synthesizes a novel view of an empty room from several object-existent room data, proving the superiority of the proposed network.

7 Acknowledgements

This work was supported by Korea Evaluation Institute Of Industrial Technology (KEIT) grant funded by the Korea government(MOTIE) (No.20023455, Development of Cooperate Mapping, Environment Recognition and Autonomous Driving Technology for Multi Mobile Robots Operating in Large-scale Indoor Workspace)

References

[1] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24, 2009.

[2] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021.

[3] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022.

[4] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5559–5568, 2021.

[5] 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.

[6] Antonio Criminisi, Patrick Perez, and Kentaro Toyama. Object removal by exemplar-based inpainting. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–II. IEEE, 2003.

[7] 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.

[8] James Hays and Alexei A Efros. Scene completion using millions of photographs. *ACM Transactions on Graphics (ToG)*, 26(3):4–es, 2007.

[9] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010.

[10] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):1–14, 2017.

[11] Mohammad Mahdi Johari, Yann Lepoittevin, and François Fleuret. Geonerf: Generalizing nerf with geometry priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18365–18375, 2022.

[12] Prakhar Kulshreshtha, Nektarios Lianos, Brian Pugh, and Salma Jiddi. Layout aware inpainting for automated furniture removal in indoor scenes. In *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 839–844. IEEE, 2022.

[13] Chen-Yu Lee, Vijay Badrinarayanan, Tomasz Malisiewicz, and Andrew Rabinovich. Roomnet: End-to-end room layout estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 4865–4874, 2017.

[14] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10758–10768, 2022.

[15] Hao-Kang Liu, I Shen, Bing-Yu Chen, et al. Nerf-in: Free-form nerf inpainting with rgb-d priors. *arXiv preprint arXiv:2206.04901*, 2022.

[16] Hongyu Liu, Bin Jiang, Yi Xiao, and Chao Yang. Coherent semantic attention for image inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4170–4179, 2019.

[17] Steven Liu, Xiuming Zhang, Zhoutong Zhang, Richard Zhang, Jun-Yan Zhu, and Bryan Russell. Editing conditional radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5773–5783, 2021.

[18] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019.

[19] 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. *Communications of the ACM*, 65(1):99–106, 2021.

[20] Ashkan Mirzaei, Tristan Aumentado-Armstrong, Konstantinos G Derpanis, Jonathan Kelly, Marcus A Brubaker, Igor Gilitschenski, and Alex Levinshtein. Spin-nerf: Multiview segmentation and perceptual inpainting with neural radiance fields. *arXiv preprint arXiv:2211.12254*, 2022.

[21] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022.

[22] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[23] Julien Philip and George Drettakis. Plane-based multi-view inpainting for image-based rendering in large scenes. In *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, pages 1–11, 2018.

[24] Konstantinos Rematas, Andrew Liu, Pratul P Srinivasan, Jonathan T Barron, Andrea Tagliasacchi, Thomas Funkhouser, and Vittorio Ferrari. Urban radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12932–12942, 2022.

[25] 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.

[26] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016.

[27] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

[28] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022.

[29] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8248–8258, 2022.

[30] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[31] Silvan Weder, Guillermo Garcia-Hernando, Aron Monszpart, Marc Pollefeys, Gabriel Brostow, Michael Firman, and Sara Vicente. Removing objects from neural radiance fields. *arXiv preprint arXiv:2212.11966*, 2022.

[32] Qianyi Wu, Xian Liu, Yuedong Chen, Kejie Li, Chuanxia Zheng, Jianfei Cai, and Jianmin Zheng. Object-compositional neural implicit surfaces. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII*, pages 197–213. Springer, 2022.

[33] Zhaoyi Yan, Xiaoming Li, Mu Li, Wangmeng Zuo, and Shiguang Shan. Shift-net: Image inpainting via deep feature rearrangement. In *Proceedings of the European conference on computer vision (ECCV)*, pages 1–17, 2018.

[34] Bangbang Yang, Yinda Zhang, Yinghao Xu, Yijin Li, Han Zhou, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. Learning object-compositional neural radiance field for editable scene rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13779–13788, 2021.

[35] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018.

[36] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Learning pyramid-context encoder network for high-quality image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1486–1494, 2019.

[37] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020.

[38] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.