Neural Rendering in a Room: Amodal 3D Understanding and Free-Viewpoint Rendering for the Closed Scene Composed of Pre-Captured Objects

BANGBANG YANG, State Key Lab of CAD&CG, Zhejiang University, China
YINDA ZHANG, Google, USA
YIJIN LI, State Key Lab of CAD&CG, Zhejiang University, China
ZHAOPENG CUI∗, State Key Lab of CAD&CG, Zhejiang University, China
SEAN FANELLO, Google, USA
HUJUN BAO, State Key Lab of CAD&CG, Zhejiang University, China
GUOFENG ZHANG∗, State Key Lab of CAD&CG, Zhejiang University, China

We, as human beings, can understand and picture a familiar scene from arbitrary viewpoints given a single image, whereas this is still a grand challenge for computers. We hereby present a novel solution to mimic such human perception capability based on a new paradigm of amodal 3D scene understanding with neural rendering for a closed scene. Specifically, we first learn the prior knowledge of the objects in a closed scene via an offline stage, which facilitates an online stage to understand the room with unseen furniture arrangement. During the online stage, given a panoramic image of the scene in different layouts, we utilize a holistic neural-rendering-based optimization framework to efficiently estimate the correct 3D scene layout and deliver realistic free-viewpoint rendering. In order to handle the domain gap between the offline and online stage, our method exploits compositional neural rendering techniques for data augmentation in the offline training. The experiments on both synthetic and real datasets demonstrate that our two-stage design achieves robust 3D scene understanding and outperforms competing methods by a large margin, and we also show that our realistic free-viewpoint rendering enables various applications, including scene touring and editing. Code and data are available on the project webpage: https://zju3dv.github.io/nr_in_a_room/.

ACM Reference Format:

Additional Key Words and Phrases: Neural rendering, 3D scene understanding, Amodal perception

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

Permission held by the owner/author(s).

© 2022 ACM, Inc.
https://doi.org/10.1145/3528223.3530163

0730-0301/2022/7-ART101

1 Introduction

Given a photo of our living room, as human beings, we can vividly picture the whole layout in our mind, including how the furniture is placed in 3D space and how the environment looks from any viewpoint, even when objects are re-arranged differently in the room. Granting the computer similar skills would require reliable indoor scene 3D semantic understanding and free-viewpoint rendering capabilities, which ideally are all fulfilled from widely available input, e.g., a single photo. Over decades, enormous efforts have been made in the field of computer vision and graphics [Dai et al. 2020; Gortler et al. 1996; Levoy and Hanrahan 1996; Mildenhall et al. 2020; Nie et al. 2020; Zhang et al. 2021a], yet the gap with the human perception is still huge. Despite this, we argue that likely humans are better at this task for places they are familiar with, and the learned prior knowledge on the objects and their arrangement in a closed room are the key to the success.

In this paper, we present a novel solution for reliable 3D indoor scene understanding and free-viewpoint rendering in a closed scene – a.k.a. a room with a fixed set of pre-captured objects but placed under unknown arrangements and diverse illuminations. Inspired by human amodal perception, our method takes advantage of an offline stage to collect prior knowledge of the target scene, where models for each object, e.g., for localization or neural rendering, can be built with an affordable workload and then fine-tuned in the specific scenario for better performance. With the help of this strong prior knowledge, during the online stage, our method only needs light-weighted input, i.e., a single panoramic image taken from the scene, and can reliably recognize and localize objects in 3D space and render the scene from arbitrary camera viewpoints via amodal 3D understanding. While general scene understanding [Nie et al. 2020; Zhang et al. 2021a] makes the best effort to make predictions under unseen environments but still suffers from generalization issues, our amodal scene understanding aims at an accurate and reliable scene understanding for familiar scenes.

A flexible yet effective scene representation is critical to the considered task. Traditional representations such as textured meshes [Izadi et al. 2017; Liu et al. 2019; Waechter et al. 2014] or voxels [Kim et al. 2013; Song et al. 2017] generally have some drawbacks, e.g., limited rendering quality [Liu et al. 2019] and resolution [Song et al. 2017], requiring pre-built CAD furniture model for scene reconstruction [Izadi et al. 2017] and explicit lighting/material definitions for lighting variations [Li et al. 2020; Matusik et al. 2003], which prohibits fine-grained scene rendering and understanding. We thus choose the neural implicit representation [Mildenhall et al. 2020] as it enables geometric reconstruction with photo-realistic volumetric rendering, and it could be extended to support functionalities such as appearance variation [Martin-Brualla et al. 2021] and scene graph decomposition [Ost et al. 2021] with rendering-based optimization.

Specifically, we first build object detection and 3D pose estimation models for all the objects of interest as well as a neural rendering model for each object, including the empty room. At run-time, given a panoramic image taken from the room stuffed with pre-captured objects in a new arrangement, the scene understanding task can be achieved by 3D object detection and pose estimation, followed by an optimization via differentiable rendering using the neural rendering models. Additionally, the per-object neural rendering models can be plugged in to support full scene free-viewpoint rendering. While this framework is technically plausible, we find it suffers from several challenges as follows, which we will address in this work:

**Intensive Computation.** Neural volume rendering methods are typically computationally intensive since a tremendous number of network queries are required for points densely sampling along pixel rays, making it prohibitive for back-propagation-based optimization, like pose estimation, where the rendering needs to be done repetitively. iNeRF [Yen-Chen et al. 2021] mitigates this issue by restricting sample pixels inside the detected region of interest, which reduces the computation cost and enables the camera pose estimation with respect to a single object on a commodity-level GPU. However, this is still not practical for room-scale environments when multiple objects need to be jointly optimized in order to handle mutual occlusions or physical relations. To tackle this challenge, we learn an implicit surface model jointly with its radiance field, inspired by NeuS [Wang et al. 2021a], which allows us to perform efficient sphere tracing [Liu et al. 2020] at the early stage of the rendering, leveraging the estimated ray-to-surface distances. Points can then be sampled from regions close to the surface, and a small number of points is sufficient for the optimization. In this way, we significantly reduce the computational cost and make it feasible to finish the joint optimization with multiple objects on a single GPU in a reasonable amount of computation.

**Incorrect Physical Relationship.** Even though machine learning models are trained per-scene, they could still make obvious mistakes like breaking the physical rules and resulting in implausible novel view rendering, e.g., objects flying in the air or intersecting with walls. To solve this problem, we propose several novel physical losses and integrate a physics-based optimization into the neural-rendering-based optimization, where the conformity to prior knowledge and even pre-defined rules (e.g., a bed should attach to the wall) are jointly optimized with the photometric error between the rendered image and the observation. This significantly helps fix errors made on individual objects and improves the overall object pose accuracy, which further delivers context abides rendering.

**Domain Gap.** The lighting condition may inevitably vary in the scene, and object renderings from the models trained at offline stage may not be consistent with the environment, which will further influence the rendering-based optimization. To mitigate this, we propose to exploit compositional neural rendering to augment the training data. In particular, we augment the pre-captured data with environment maps sourced from polyhaven.com [Zaal et al. 2020], and learn the neural rendering models conditioned on lighting represented in a latent space. During the neural rendering based optimization, the neural rendering model is able to respond to novel illumination other than the one during the pre-capture stage, and both the environment lighting and object pose can be successfully optimized. We also synthesize objects with different scene layouts and render photo-realistic images for the training of object prediction, which empirically enhances model robustness.

Our contributions can be summarized as follows. We present a practical solution for a novel task which aims at amodal 3D scene understanding and free-viewpoint rendering for indoor environments from a single panoramic image. We design a two-stage framework...
in which per-object pre-trained models are learned offline, and a neural-rendering-based optimization is exploited for online 3D understanding. We analyze the technical challenges for this novel task and propose mitigation techniques to improve run-time efficiency, add physical constraints in the model, handle illumination changes, and increase the data diversity that is hard to be ensured physically. Extensive experiments show that our method can achieve significantly better 3D scene understanding performance than state-of-the-art general 3D scene understanding methods and meanwhile, deliver free-viewpoint rendering capability that supports high-level applications like scene editing and virtual touring.

2 Related Work

3D Scene Understanding. 3D scene understanding is a popular topic in computer vision. Early works mainly focus on room layout estimation with Manhattan World [Coughlan and Yuille 1999; Rama-lingam et al. 2013; Sun et al. 2019a; Yan et al. 2020; Kou et al. 2021] or cuboid assumption [Dasgupta et al. 2016; Mallya and Lazebnik 2015; Song et al. (Song et al. 2015) attempts to reconstruct and recognize scene objects from a domestic robot but requires laborious crowd-sourced annotations. With the advance of neural networks, many works propose to estimate both object poses and the room layout [Du et al. 2018; Huang et al. 2018a; Zhang et al. 2017]. To recover object shapes, some methods [Ben et al. 2019; Grouex et al. 2018; Wang et al. 2018] reconstruct meshes from a template, and others [Huang et al. 2018b; Ijadina et al. 2017] adopt shape retrieval approaches to search from a given CAD database. Recently, some approaches [Dahnert et al. 2021; Nie et al. 2020; Popov et al. 2020; Yang and Zhang 2016; Yang et al. 2019; Zhang et al. 2021b] enable 3D scene understanding by generating a room layout, camera pose, object bounding boxes, or even meshes from a single view, automatically completing and annotating scene meshes [Bokhovkin et al. 2021] or predicting object alignments and layouts [Avetisyan et al. 2020] from an RGB-D scan. Inspired by PanoContext [Zhang et al. 2014] that panoramic images contain richer context information than the perspective ones, Zhang et al. [Zhang et al. 2021a] propose a better 3D scene understanding method with panoramic captures as input. For amodal scene completion, Zhan et al. [Zhan et al. 2020] propose to decompose cluttered objects of an image into individual identities. However, these works still suffer from limited generalization in real-world environments and do not allow fine-grained scene presence from arbitrary views.

Neural Rendering. Neural rendering methods aim at synthesizing novel views of objects and scene by learning scene representation from 2D observations in various forms, such as voxels [Lombardi et al. 2019; Sitzmann et al. 2019a], point clouds [Dai et al. 2020], meshes [Riegler and Koltun 2020, 2021], multi-plane images [Mildenhall et al. 2019; Tucker and Snively 2020; Wang et al. 2021b] and implicit functions [Mildenhall et al. 2020; Niemeyer et al. 2020; Sitzmann et al. 2019b]. NeRF [Mildenhall et al. 2020] uses volume rendering to achieve photo-realistic results; follow up works extend the model to multiple tasks, such as pose estimation [Yen-Chen et al. 2021], dense surface reconstruction [Oechsle et al. 2021; Wang et al. 2021a; Yariv et al. 2021] and scene editing [Granskog et al. 2021; Guo et al. 2020; Yang et al. 2021]. Meanwhile, other methods [Riegler and Koltun 2020, 2021] also show impressive free-viewpoint rendering capability in the wild, or scene rendering [DeVries et al. 2021; Luo et al. 2020] of indoor environments. However, existing neural rendering pipelines either need to be trained for a static scene thus do not generalize to dynamic environments, or require domain prior [Wang et al. 2021b; Yu et al. 2021], limiting the free-viewpoint rendering in unconstrained settings.

3 Method

Given a panoramic image of a closed environment with unknown furniture placement, our goal is to achieve reliable 3D scene understanding, including instance semantic detection, 3D geometry of each object, and their arrangements (i.e., object positions) in the room, utilizing the data pre-captured beforehand. We split the
whole pipeline into an offline stage and an online stage. During the **offline stage**, we scan each object and the scene background with an RGB-D camera, train a neural implicit renderer for every object of interest in the room, and then fine-tune object detectors via compositional neural rendering. In the **online stage**, as shown in Fig. 2, we first predict object meta information (i.e., poses, segmentation and IDs) from the panoramic image, and then follow pre-defined rules to generate object-object and object-room relations based on the object prediction and geometric cues (e.g., physical distances obtained from the encoded neural implicit model). Finally, to correctly estimate the scene arrangement and lighting condition that visually fits the input panorama, we perform holistic optimization with all the photometric and geometric cues, which further enables free-viewpoint scene touring and scene editing.

### 3.1 Offline Stage

#### 3.1.1 Neural Implicit Renderer

**Neural Implicit Model for Scene and Objects.** We use a neural implicit renderer that hierarchically encodes the room. Practically, we choose the SDF-based implicit field for geometry representation [Wang et al. 2021a; Yariv et al. 2021]¹, since it provides an exact surface to facilitate geometric optimization, e.g., for collision detection, while NeRF’s density field is too noisy or uncertain to support a similar objective. As shown in Fig. 2, we separately express geometry in SDF values (with SDF surface model $F_{\text{SDF}}$) and colors (with radiance model $F_R$). We will show later that this formulation enables efficient neural-rendering-based optimization by providing geometric cues like ray intersection distances with sphere tracing. Motivated by Yang et al. [Yang et al. 2021], we encode scene background and objects in two branches, and use the object code $I_{\text{obj}}$ to control the visibility of a certain object, rather than per-model per-object training. We render the object $k$ with sampled points $\{x_i | i = 1, ..., N\}$ along the ray $r$, which is defined as:

$$
\hat{C}(r)_{\text{obj}} = \sum_{i=1}^{N} T_i \alpha_{\text{obj}} e_{\text{obj}} \cdot \sum_{j=1}^{i-1} (1 - \alpha_{\text{obj}}) T_j, \\
\alpha_{\text{obj}} = \max \left( \frac{\Phi_k(SDF(x_i)) - \Phi_k(SDF(x_{i+1}))}{\Phi_k(SDF(x_i))}, 0 \right).
$$

Note that we omit the object index $k$ for brevity. $T_j$ is the accumulated transmittance, $T_j$ is the logistic density distribution, $SDF(x) = F_{\text{SDF}}(x, I_{\text{obj}})$, and $\alpha_{\text{obj}}$ is the opacity value derived from the SDF surface model. $e_{\text{obj}}$ is the color defined as $e_{\text{obj}} = F_R(x, v, I_{\text{obj}}, I_l, I_p)$, where $v$ is the viewing direction, $I_l$ is the appearance code [Martin-Brualla et al. 2021] that handles per-frame sensing variations (e.g., white balance and auto-exposure on real-world data), and $I_p$ is the light code introduced later. We supervise the renderer with color, depth and object masks, and jointly render multi-objects and scene background by ordering the distance of samples along the ray directions and render pixels $\hat{C}(r)$ following the quadrature rules. More details can be found in the supplementary material.

**Lighting Augmentation & Light Code Learning.** We learn a neural renderer conditioned on a latent lighting space $I_l$, aiming at modeling scene-level illumination variation and adapting to the target scene depicted in the given panorama. Since it is non-trivial to capture real-world images with thorough lighting variation, we synthetically augment the captured image with diffuse shading rendered from realistic HDR environment maps. Practically, we gather 100 HDRI indoor environment maps from the Internet [Zaal et al. 2020] and convolve them to diffuse irradiance maps [Debevec 2006]. Then we compute the per-pixel surface normal in the world coordinate and retrieve the corresponding light intensity from the light map. Finally, we multiply the light intensity to the input images. However, for real-world data, reliable surface normals are not readily available. To tackle this problem, we leverage a two-stage pipeline by first training a naïve neural renderer without augmentation, and then extracting mesh from the model for normal computation. We show this procedure in Fig. 3, where the captured image has been naturally augmented with two different light maps. During the training stage, we randomly augment the input images with pre-convolved light maps and feed the radiance model $F_R$ with a learnable light code $I_l^m$ where $m$ is the index of the light map. Although such geometry-aware augmentation does not cover every important aspect of the real-world physics and provides augmented data only up to an approximation, it brings convenience to the offline stage: the training data is collected only once under a mild lighting condition, and the network empirically adapts to unseen lighting decently (see Sec. 4.3).

#### 3.1.2 3D Object Prediction Fine-tuning

**Module Design.** As illustrated in Fig. 2 (on top left), we adopt the object detector (ODN) from Zhang et al. [Zhang et al. 2021a, 2014] and Nie et al. [Nie et al. 2020] to detect scene objects and estimate object poses w.r.t. the camera, and use an object identifier based on NetVLAD feature similarity [Arandjelovic et al. 2016] to recognize previously seen objects.

**Data Augmentation with Neural Rendering.** At training time for each scene, instead of physically moving objects in the real-world, we exploit compositional neural rendering with implicit

---

¹In our paper, we use the formulation from NeuS [Wang et al. 2021a], but VoSDF [Yariv et al. 2021] is also applicable.
neural renderer (Sec. 3.1.1) to render labeled panorama for training, where objects are randomly placed following user-defined rules (e.g., bed and table should attach to the floor). Then, we perform a fast fine-tuning for the pre-trained ODN network from Zhang et al. [Zhang et al. 2021a] and also store the NetVLAD features for each object of different views.

3.2 Online Stage

3.2.1 Bottom-up Initialization

**Object Prediction.** We first feed panoramic images to the object detector, and obtain object meta information including initial pose estimation, instance / semantic segmentation and object identities.

**Relation Generation.** As demonstrated in [Nie et al. 2020; Zhang et al. 2021a, 2014], indoor scenes are commonly well-structured (e.g., beds and nightstands are usually attached to the wall, desks and chairs are often supported by the floor), and such prior knowledge can be formulated as various relations to enhance arrangement optimization. Therefore, we also generate a series of relations for physical constraints optimization (Sec. 3.2.4), including object-object support, object-wall attachment and object-floor support. Practically, we directly infer relations based on object meta information and geometric cues (extracted bounding boxes, ray intersection distance and normal) from the SDF surface model with user-defined rules (see supplementary material). In theory, our method can also work with rules or scene context learned in a data-driven way [Zhang et al. 2021a], which we leave for future work.

**Camera Pose Estimation.** Since the optimization is based on a known neural implicit model, we need to locate camera poses to ensure background rendering is aligned with the input image. To do so, we transform the panorama to multiple perspective views (i.e., similar to “equirectangular to cubemap” conversion by warping pixels according to ray directions) and employ the method from Sarlin et al. [Sarlin et al. 2019, 2020] for visual localization.

**Object Pose Parameterization.** We optimize poses $\hat{\mathbf{T}}^k \in \text{SE}(3)$ for $K$ objects, where the rotation $\hat{\mathbf{R}}^k$ is parameterized as Zhou et al. [Zhou et al. 2019]., and the position (a.k.a. object center) $\hat{\mathbf{p}}^k$ is directly expressed in Euclidean space.

3.2.2 Photometric Constraint Optimization

**Tone Adjuster.** To better adapt the lighting condition to the input panorama at the online stage, we introduce a per-object tone adjuster which explicitly models lighting variations and helps to reduce the burden of light code optimization. In practice, we additionally optimize a learnable shifting factor $t_{\text{obj}}^k$ and scaling factor $s_{\text{obj}}^k$ for each object $j$ as: $\hat{c}_{\text{obj}}^k = (c_{\text{obj}}^k - t_{\text{obj}}^k) s_{\text{obj}}^k + t_{\text{obj}}^k$, which can be regarded as color transformation [Reinhard et al. 2001] but in a per-object manner. We find this explicit representation benefits the lighting adaptation, as demonstrated in our experiments.

**Photometric Loss with Joint Rendering.** We use photometric constraint by leveraging joint rendering where the photometric loss is back-propagated to optimize per-object poses and light parameters. For each input image, we sample $N$ rays on the object masks and $0.2N$ rays on the background so as to ensure the convergence of both objects and background. The photometric loss is defined as the squared distance between rendered colors $\hat{C}(r)$ and pixel colors $C(r)$ from the input panorama for all the sampled rays $r \in N_r$:

$$L_{\text{pho}} = \frac{1}{|N_r|} \sum_{r \in N_r} ||\hat{C}(r) - C(r)||^2_2.$$ (2)

**Safe-Region Volume Rendering.** However, neural volume rendering requires hundreds of network queries for each ray, which restricts tasks like pose estimation [Yen-Chen et al. 2021] by only sampling a small bunch of rays due to the limitation of GPU memory. This is particularly true in our task as one ray might go through 2 or 3 objects at a time when object to object occlusions happen, which results in 2 or 3 times more queries than a single object case. Fortunately, as our renderer learns an SDF-based representation of the geometry, we can easily determine ray intersections using sphere tracing at the early stage of the rendering. Inspired by Liu et al. [Liu et al. 2020], we propose a safe-region volume rendering by first computing ray-to-surface distances with efficient sphere tracing and then sampling much fewer points near the surface for differentiable volume rendering. Our experiments demonstrate that this strategy significantly reduces network query times and allows us to jointly optimize more objects in cluttered scenes. Please refer to the supplementary material for more details.

3.2.3 Observation Constraint Optimization

**Observation Loss.** The initial poses from object prediction may be inaccurate on the dimension of camera-to-object distance due to scale ambiguity, but the observing angles (a.k.a. object center re-projection) on the panoramic view are usually reliable. Thus, we also add an observation constraint by encouraging closer observing angles of objects between initial pose estimation and the optimized pose, as:

$$I_{\text{obs}} = \sum_{k=1}^{K} ||1 - \text{sim}(\hat{p}_\text{init}^k - p_{\text{cam}}, \hat{p}_k - p_{\text{cam}})||^2,$$ (3)

where $\text{sim}(\cdot)$ denotes cosine similarity, and $p_{\text{cam}}$ is the camera center estimated in Sec. 3.2.1.

3.2.4 Physical Constraint Optimization

Prior scene understanding works [Nie et al. 2020; Zhang et al. 2021a, b] mainly build physical constraints upon object bounding boxes and room layout under Manhattan assumption. Thanks to the precise geometries encoded in the neural SDF model, we can define physical constraints to optimize physical conformity at a finer-grained level.

**Magnetic Loss.** We introduce a novel magnetic loss that fully leverages neural renderer’s SDF model to optimize generated relations (e.g., attachment and support) from Sec.3.2.1. As the name suggests, the magnetic loss encourages two opposite surfaces of the attached objects to be close to each other without violation. Practically, we shoot a set of probing rays from one object surface plane to another with the shooting direction from the generated relations (Sec. 3.2.1) and compute ray-to-surface distances via sphere tracing. Then, as illustrated in Fig. 4, we define two distances to diagnose surface-surface relations: 1) attachment distance $d_a$ which measures the surface distance between two objects by summing up the distances of partial nearest intersections while ignoring far-away intersections, 2) violation distance $d_v$ which indicates potential violation of
two objects by summing up all the violation part of the surfaces. To this end, we defined magnetic loss as:

\[ L_{\text{mag}} = \frac{1}{K} \sum_{k=1}^{K} \max(d^k_{\text{att}}, 0) + \max(d^k_{\text{vio}}, 0). \]

Please refer to the supplementary materials for more details.

**Physical Violation Loss.** To mitigate physical occlusions that the magnetic loss does not cover (e.g., chair under the desk), inspired by Zhang et al. [Zhang et al. 2021b], we add a physical violation loss based on the neural SDF model. Different from Zhang et al. [Zhang et al. 2021b] which uniformly samples points inside the bounding boxes, we only sample points on the visible surface with outside-in sphere tracing, so as to make the optimization more efficient when two objects only collide partially at a shallow level. The physical violation loss is defined by punishing \( P \) surface points for each object \( k \) when an object’s points lie inside its \( O \) neighbor objects by querying the corresponding SDF surface model \( F_{\text{SDF}} \) as:

\[ L_{\text{vio}} = \sum_{k=1}^{K} \sum_{o=0}^{O} \sum_{p=1}^{P} \min(\text{SDF}(x_{kp})_o + \epsilon, 0), \]

where \( o = 0 \) denotes the scene background, and we set \( \epsilon = 0.025 \), \( O = 3 \) and \( P = 1000 \) in our experiment.

**Gravity Direction Loss.** In real-world scenarios, many furniture like beds and tables only rotate around the gravity direction (i.e., rotation uncertainty only on the yaw angle). So we also add the gravity-direction energy term to the physical constraints for those objects as:

\[ L_g = \sum_{j=1}^{K} \text{sim}(\hat{r}^k, g), \]

where \( g = [0, 0, 1]^T \) is the gravity direction.

The overall physical loss is defined as:

\[ L_{\text{phy}} = L_{\text{mag}} + L_{\text{vio}} + L_g. \]

### 3.2.5 Holistic Optimization

In holistic optimization, we seek for per-object poses \( \hat{T}^k \), object and background appearance codes \( l^k_o \) and light codes \( l^k_l \) that satisfy the input panoramic image. To fulfill this goal, we jointly optimize photometric loss, observation loss and physical constraint losses at the online stage, as:

\[ L = \lambda_{\text{pho}} L_{\text{pho}} + \lambda_{\text{obs}} L_{\text{obs}} + \lambda_{\text{phy}} L_{\text{phy}}. \]

We use \( \lambda_{\text{pho}} = 1, \lambda_{\text{obs}} = 100, \) and \( \lambda_{\text{phy}} = 1 \) in our experiment. The total optimization takes about 10-15 minutes (depending on the frequency of object occlusions) for a panoramic image with 500 iterations on an Nvidia RTX3090-24G graphics card. More discussion of the time-consuming and the possible improvement at the online stage can be found in Sec. 5.

### 4 Experiments

In this section, we first compare our scene arrangement prediction with DeepPanoContext [Zhang et al. 2021a] and evaluate the scene lighting adaptation ability both quantitatively and qualitatively. Then, we perform ablation studies to analyze the design of our framework. Finally, we demonstrate the applicability of our method on scene touring, scene illumination interpolation, and scene editing.

#### 4.1 Dataset

**iG-Synthetic.** We use iGibson [Shen et al. 2020] simulator to synthesize labeled images with depth, segmentation and 3D bounding boxes for training and testing. For training object-centric models for identification, pose estimation or neural rendering, we generate 360° views around each object (similar to Realistic Synthetic 360° in NeRF [Mildenhall et al. 2020]). For the background scenes, we leverage the toolbox from Zhang et al. [Zhang et al. 2021a] to generate panoramic views of the iGibson scenes. Since many rooms in iGibson are either too empty (e.g., bathroom and storage-room) or filled with fixed stuff (e.g., basin and oven in kitchen), we thus select four representative scenes (i.e., bedroom, lobby, child’s room and home office) which already covers most of the movable object types in the dataset.

**Fresh-Room.** To demonstrate the efficacy in real-world scenes, we create a new dataset named Fresh-Room, which contains RGB-D posed training images for 5 objects and the room background captured by iPad Pro. We also capture multiple panoramic testing images under 4 different setups with varying arrangements and lighting conditions using a 360° camera (Insta360 ONE-R). We utilize the SFM system with mesh reconstruction [Kazhdan et al. 2006; Schönberger and Frahm 2016] and ARKit metadata2 to recover camera poses with real-world scale and obtain 2D segmentation by projecting annotated labels from 3D meshes for training data.

#### 4.2 Scene Arrangement Prediction

We first evaluate scene arrangement prediction on iG-Synthetic dataset and our Fresh-Room dataset. For iG-Synthetic dataset, we reorganize the scene arrangement following room examples [Shen et al. 2020], producing unseen arrangements for four scenes, and synthesizing testing data with 5 unseen indoor illuminations based on the iGibson PBR engine. Since there is no amodal scene understanding approach for comparison, we take DeepPanoContext [Zhang et al. 2021a] (Pano3D) as a reference, which is a SOTA method for general holistic 3D scene understanding with a panoramic input. The Intersection over Union (IoU), Average Rotation Error (ARE) and Average Position Error (APE) are used as evaluation metrics. As demonstrated in Fig. 5 and Tab. 1, our method consistently achieves
Fig. 5. Scene arrangement visualization with textured meshes, where objects are scaled by the estimated bounding boxes. Note that for the Fresh-Room, we use the object meshes extracted from our neural implicit renderer.

### Table 1. Quantitative evaluation on scene arrangement prediction.

<table>
<thead>
<tr>
<th>Scene</th>
<th>DeepPanoContext</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parkinson</td>
<td>IoU (%)</td>
<td>ARE (°)</td>
</tr>
<tr>
<td>Bedroom</td>
<td>26.34</td>
<td>52.80</td>
</tr>
<tr>
<td>Child’s Room</td>
<td>40.61</td>
<td>30.33</td>
</tr>
<tr>
<td>Home Office</td>
<td>27.76</td>
<td>34.64</td>
</tr>
<tr>
<td>Average</td>
<td>33.19</td>
<td>36.31</td>
</tr>
</tbody>
</table>

### 4.3 Scene Lighting Adaptation

Since we decouple lighting variation implicitly in a latent space ($l_l$) and explicitly via tone adjuster, hence we can adjust the renderer to fit the lighting condition at test time. As shown in Fig. 6, the input images (first row) have dramatically different lighting variations such as local highlight and global warm light. When the lighting adaptation is disabled (second row), the rendered results are close to the pre-captured training views, where the rendered furniture comes up with inconsistent shininess (e.g., the shelf in the third column are inevitably brighter than the real image). By introducing implicit light code optimization (third row), the rendered scenes are closer to the input ground-truth but struggles to adapt to the extremely warm light in the last column, where the global tone has turned yellow but the floor color and the curtain color are distorted. By enabling the tone adjuster (fourth row) only, we can also handle a certain degree of lighting variation (e.g., carpet and desk lit by yellow light in the second row, warm light in fourth column), but fails to adapt the local lighting variation (e.g., scene background partially lit by strong light in first column). When the tone adjuster and the light code optimization are both enabled, we successfully render images with local highlight and global consistent tone, and also achieve the best metric performance as demonstrated in Tab. 2. We believe that the tone adjuster effectively reduces the burden of latent space optimization, and the combination of explicit and implicit optimization enhances the lighting adaptation ability.

### 4.4 Ablation Studies

#### Data Augmentation. We first analyze the effectiveness of our data augmentation with compositional neural rendering for object prediction. Specifically, we use the ODN network [Nie et al. 2020; Zhang et al. 2021a] as a baseline and fine-tune on the labeled panoramic images rendered by our neural implicit renderer. The results in Tab. 3 show that our proposed data augmentation improves the object prediction quality on the rotation error and IoU (first two rows), and also boosts the performance of scene arrangement prediction in holistic optimization (third row and last row).  

#### Various Constraints. We then inspect the efficacy of various constraints in our holistic optimization, including photometric constraint in Sec. 3.2.2, observation constraint in Sec. 3.2.3 and physical constraint in Sec. 3.2.4. Note that we exclude lighting adaptation as this process is mainly for better rendering quality and performed after the object pose optimization (see supplementary for more details). As shown in Tab. 3 (last five rows), all these loss terms improve the overall scene arrangement quality. Furthermore, we evaluate the performance when only a bare photometric loss is enabled. It is clear that without physical and observation constraints, a simple photometric loss is prone to be unstable in such cluttered scenes. 

#### Safe-Region Volume Rendering. We also compare our proposed safe-region volume rendering with the classical volume rendering
pipeline (e.g., used by iNeRF [Yen-Chen et al. 2021]) in the pose optimization task with an Nvidia RTX-3090-24GB graphics card. During the experiment, we optimize object poses by jointly rendering background scenes and target objects with back-propagation, and report the GPU memory usage by varying the number of sample rays and target objects. As shown in Tab. 4, our proposed strategy significantly reduces the number of network queries and GPU memory consumption and can simultaneously optimize 10 objects, while the classical volume rendering fails due to out of memory. We verify the impact on the pose estimation quality in the supplementary.

4.5 Free-viewpoint Scene Touring
Once we resolve the scene arrangement and scene lighting condition, it is feasible to re-render the room in any arbitrary view, which enables virtual touring of the room. To inspect the rendering quality for this task, we conduct a scene re-rendering experiment by fitting input images (first column in Fig. 7) and render another view with the fitting results. Thanks to our neural scene representation, the rendered novel views (last column in Fig. 7) vividly reproduce scene appearance and lighting conditions (e.g., local highlight and global warm light) of the corresponding ground-truth novel views (second column in Fig. 7). As a comparison, when ablating the lighting adaptation from the representation, we can still achieve realistic novel view rendering results, but the specific lighting conditions (e.g., local highlight and warm tone) are no longer kept (third column in Fig. 7), which also results in lower metric performances.

4.6 Scene Editing & Illumination Interpolation
Since our neural implicit renderer has already learned to render the scene background and objects, we can easily edit or composite novel scenes upon this. As shown in Fig. 8, we perform scene editing by inserting virtual objects learned from iG-Synthetic into the real scene Fresh-Room, and the rendered image of novel view demonstrates correct space relationship with seamless object-object occlusion. We also conduct the illumination interpolation experiment in Fig. 9, where the scene lighting temperature can be naturally turned from day to night (first row), or from cold to warm (second row).

5 Conclusion
We propose a novel 3D scene understanding and rendering paradigm for closed environments. Given a single panoramic image as input, our method can reliably estimate 3D scene arrangement and the lighting condition via a holistic neural-rendering-based optimization framework. The proposed method also enables free-viewpoint scene touring and editing by changing illumination or objects’ placement. Despite the novel capabilities provided by our method, it still has its limitations. First, since we assume the neural implicit models are pre-built, our method cannot handle the cases with unobserved
Acknowledgments

We thank Hanqing Jiang and Liyang Zhou for their kind help in object reconstruction. This work was partially supported by NSF of China (No. 62102356, No. 61932003) and Zhejiang Lab (2021PE0AC01).

References


Supplementary Material

In this supplementary material, we will describe more details of our framework including object prediction in Sec. A, neural implicit renderer in Sec. B, and holistic optimization in Sec. C. Besides, we also provide more discussion in Sec. D and more experimental results in Sec. E.

A Details of Object Prediction

A.1 Object Detection

Implementation Details. We adopted the Object Detection Network (ODN) from [Nie et al. 2020; Zhang et al. 2021a] to predict the initial object poses. The vanilla ODN extracts the appearance feature with ResNet-18 from 2D detection, and encodes the relative position and size between 2D object boxes into geometry features. Differently, to generalize better with our data augmentation where the objects are randomly placed in the neural rendered views, we replace the object relative features with object sizes and positions of themselves. Even though we do not rely on object relation knowledge, we still obtain performance gains in the object pose estimation thanks to the data augmentation.

Training and Testing Details. The ODN is fine-tuned on the labeled panoramic images synthesized by our neural implicit renderer. We train the network for 10 epochs with a batch size of 6 in one hour on a single Nvidia RTX-3090-24G graphics card. During the testing stage of the Fresh-Room dataset, we also pre-process the input image with automatic white-balance correction [Afifi et al. 2019] to reduce the impact of lighting variation on 2D detection, which is also applied to DeepPanoContext [Zhang et al. 2021a] for fair comparison. Moreover, according to the statistics, the testing data of iGibson-Synthetic contains frequent occlusions (with about 16% in Lobby, 35% in Bedroom, 31% in Child’s room, and 40% in Office room).

A.2 Object Identification (ReID)

During the ODN training phase, we generate extensive labeled panoramas by compositional neural rendering with the implicit neural renderer. Each 2D object boxes in the equirectangular images is represented as a Bounding FoV (BFoV) [Yang et al. 2018], which is defined with the latitude and longitude of the center and the horizontal and vertical field of view. We crop out the objects in each BFoV from the panoramas, warp them into perspectives and then save the NetVLAD features of these perspectives with their corresponding object ids as our retrieval database.

In the identification phase, we compute the NetVLAD features of each 2D detection boxes, and then find the one with the highest similarity in the database. We remove the query boxes with a match similarity threshold less than $s = 0.25$. The remaining query boxes are identified as the most similar objects in the database.

B Details of Neural Implicit Renderer

B.1 Model Architecture

We adopt NeuS [Wang et al. 2021a] as the network backbone for the neural implicit renderer. As introduced in Sec. 3.1.1, we append 64-dimension object codes $I_{obj}$ to the input of the SDF surface model $F_{SDF}$, so as to control the visibility of a certain object identity. For the radiance model $F_{R}$, we add 16-dimension appearance code $I_{a}$ to handle the sensing variation on the Fresh-Room dataset, and add 16-dimension light code $I_{l}$ to learn the latent space of illumination variation. Besides, to enhance the rendering texture details on the Fresh-Room dataset, we also extend the positional encoding [Mildenhall et al. 2020] of the spatial coordinate to 10 in $F_{R}$.

B.2 Training Details

Following Yang et al. [Yang et al. 2021], we use color, depth, and object losses to supervise the neural implicit renderer. To adapt the geometric initialization in the SDF surface model to the inside-out reconstruction for room background, we reverse the sign of the bias in the last MLP [Gropp et al. 2020] during network initialization, so the initial surface normal is facing towards the scene center. Due to the limitation of network capacity, we do not learn every objects in one MLP, but encode 4 or 5 objects per model for the iG-Synthetic dataset and 1 object per model for the Fresh-Room dataset. Note that we use a different model setting for real/synthetic data, which is a trade-off between model capacity, training time and rendering quality. Since the appearance detail of Fresh-Room is richer than iGibson, we thus deliver the best visual quality with a per-model-per-object setting. Please refer to the Fig. J for the visualization of rendering quality of the Fresh-Room dataset under different model settings. During training, we adopt the Adam optimizer with an initial learning rate of 0.0005 and a polynomial decay schedule with the power of 2. The training process for each neural model takes about 10 hours with a batch size of 1024 rays on a single Nvidia RTX-3090-24G graphics card.

C Details of Holistic Optimization

C.1 Details of Relation Generation

As introduced in Sec. 3.2.1, we use the object meta information from object prediction and geometric cues (extracted bounding boxes,
We use a safe-region volume rendering strategy in photometric constraint optimization, including object-object support, object-wall attachment and object-floor support. In practice, we put the object with the initial pose prediction and detect surface-to-surface distances with a set of probing rays (similar to that in magnetic loss) along with several pre-defined directions (downward for support relations, and four perpendicular directions to the bounding box surfaces for attachment relation) and find the relation candidates with distances smaller than 0.5m. For the attachment relation candidates, we filter the case that the surface is not opposite to the walls by ensuring intersection point normal almost opposite to the probing ray shooting directions, e.g., we only allow the back face of a cabinet to attach to the wall rather than the front face. Finally, we apply custom rules to further filter out unexpected relations and ensure several definite relations (e.g., bed, shelf, nightstand and wall-pictures are always attach to wall; bed, chairs and bottom cabinet are always attach to the floor; PC monitor, wall-pictures and mirrors should not attach to the floor). During the holistic optimization, we periodically regenerate relations per 50 iterations, so the potential relations that cannot be found at the beginning can be supplemented when object poses are better converged.

C.2 Details of Photometric Constraint Optimization
We empirically sample 1024 rays with the resolution of 320×160 when optimizing photometric loss. Because the equirectangular images are squeezed on vertical center and stretched on top/bottom poles, we draw a biased sampling with more rays near the vertical center and less when close to the poles. For the ray samples on the object masks, we adopt the region-based sampling [Yen-Chen et al. 2021] by dilating object masks with 5×5 kernels. In practice, we first optimize object poses and then optimize appearance and light codes (with zero-initialized) while freezing object poses, since we found the convergence of poses and lighting is more stable in this way.

C.3 Details of Safe-Region Volume Rendering
We use a safe-region volume rendering strategy in photometric constraint optimization, where we only sample a few points near the surface in volume rendering. Practically, for each ray, we first perform an efficient sphere tracing to find the ray-intersection point, where the ray-to-surface distance is denoted as $d_{\text{surf}}$. Then, we empirically bound NeRF’s hierarchical sampling near-far range to $d_{\text{surf}} = 0.05$ and $d_{\text{surf}} + 0.1$, and only sample 8 points for “coarse” sampling and 16 points for “fine” sampling. Besides, we also analysis the impact to the pose convergence w.r.t. the number of sample points of this strategy in Sec. E.2.

C.4 Details of Magnetic Loss
We propose magnetic loss in Sec. 3.2.1, which faithfully attach two opposite surfaces for two objects that have been marked as attachment or support relations. As explained in the main paper, we use several probing rays to determine attachment distance $d_a$ and violation distance $d_v$. In our experiment, to handle the potential violations that hinders sphere tracing (e.g., objects inside the wall), the starting points of the probing rays have always been back-offset for a certain distance $d_{\text{back}}$. The $d_{\text{back}}$ is starting from 0.5m, and is gradually increased until the sphere-tracing finds the intersection or $d_{\text{back}}$ exceeds the preset upper-bound (1.5m). For the surface plane that the ray array comes from, we use a similar ray-intersection test to find the surface points for attachment relations (e.g., looking for the side of the bed and ignoring the hollow part). Besides, for support relations, we directly use extracted bounding box bottom plane (e.g., using bed’s bounding box bottom as surface plane, since the bed legs or cabinet legs might be too tiny for the probing-ray-based intersection detection).

D More Discussions
D.1 Applicability for Real-World Scenes
When running with complex real-world scenes, since we use SDF-based constraints instead of bounding boxes, the holistic optimization also supports cluttered objects (with gravity loss omitted for certain object types). However, we need to refactor the object detector [Sun et al. 2019b; Tan et al. 2020] and relation generator [Xu et al. 2019, 2022] to support cluttered objects like books and pens. Besides, for the rooms containing windows with dynamic landscapes, we recommend applying masks to the images. And it might also be possible to simplify the offline stage (e.g., using domestic robots instead of hand-held capturing) when deploying to real applications.

D.2 Scale with Number of Objects
The computation resources (e.g., GPU memory) of our method grow linearly with the number of different objects, and we’ve tested 13 objects in the experiments. Besides, for the scene with massive similar objects (e.g., a classroom scene with 20-50 similar desks), we can reuse rendering models to save memory.

D.3 Offline-Stage Efforts
Generally, we take 30 minutes to capture the Fresh-Room data, and spend 15 minutes to annotate all the object meshes. The other process (e.g., reconstruction and training) can be automated and trained in parallel within 20 hours. Note this needs to be run only once.

D.4 Objects with the Same Shape but Different Colors
Sometimes, we may have two objects with identical shape but just different colors (e.g., a red chair and a blue chair). Note that our method does not handle such cases, as this is not available in our dataset. In the future, we can disentangle color and geometry by simply enforcing shape coherence (e.g., Edit-NeRF [Liu et al. 2021]) to benefit applications when handling mass-produced furniture like IKEA.
Table E. ReID Performance on the iG-Synthetic dataset.

<table>
<thead>
<tr>
<th>Scene</th>
<th>Lobby</th>
<th>Bedroom</th>
<th>Child’s Room</th>
<th>Home Office</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>91.84</td>
<td>78.79</td>
<td>84.48</td>
<td>88.72</td>
<td>85.96</td>
</tr>
<tr>
<td>Recall</td>
<td>86.33</td>
<td>76.42</td>
<td>75.00</td>
<td>79.09</td>
<td>79.21</td>
</tr>
</tbody>
</table>

Fig. K. Analysis of safe-region volume rendering on pose convergence performance. S.R.8-16 denotes safe-region volume rendering with 8 "coarse" samples and 16 "fine" samples. w/o S.R denotes classical volume rendering with 64 "coarse" samples and 64 "fine" samples.

E More Experimental Results

E.1 Performance of Re-Identification

We evaluate the performance of our ReID module on the iG-Synthetic dataset with four scenes (bedroom, lobby, child’s room and home office). The scenes are re-organized to unseen arrangements and unseen indoor illuminations, as described in the original paper. We adopt two commonly-used metrics, i.e., precision and recall, to measure the ReID performance, as shown in Table E.

Although the panoramic image can capture all the unoccluded objects, there are still some objects that are too far away or severely squeezed due to equirectangular projection, which are hard to be detected and affect the performance of ReID. These missing detection can be handled by introducing positional prior in our future work, where we can assume the furniture arrangement is only gradually changing for daily lives.

E.2 Analysis of Safe-Region Volume Rendering

We analyze the impact of safe-region volume rendering on the pose convergence w.r.t. the number of sample points. The experiment is conduct by optimizing bed poses from the iG-Synthetic bedroom with 1024 ray samples, so the classical volume rendering is also capable to optimize without out-of-memory error. As shown in Fig. K, even when only sample one fourth of the points (S.R.8-8) along each ray, we can still achieve similar convergence performance both on position error and rotation error.

E.3 Scene Arrangement

We show more qualitative comparisons with DeepPanoContext (Pano3D) [Zhang et al. 2021a] in Fig. L, Fig. M, and Fig. N (extended bedroom scene of Fresh-Room dataset). It is clear that DeepPanoContext can not precisely estimate object poses (e.g., the bed and the bottom cabinet are tilted in iG-Synthetic Bedroom, see Fig. L), and the estimated size are distorted in the Fresh-Room dataset (see Fig. M and Fig. N). In contrast, as our neural-rendering-based holistic optimization takes both photometric and physical constraint into consideration, our method successfully estimates object arrangement with correct poses and accurate scales.

E.4 More Examples on Lighting Augmentation

We show more examples of lighting augmentation in Fig. O and Fig. P, where the scene background and objects have been augmented with different lighting conditions.

F Social Impact

Our method may suffer from potential privacy risks, since it involves recording sensitive information of the scene. To mitigate this, the user should be notified in advance and real systems should ensure that the data does not leave the users’ device and could be deleted upon request.
Fig. L. More examples of Scene Arrangement on the iG-Synthetic dataset.
Neural Rendering in a Room: Amodal 3D Understanding and Free-Viewpoint Rendering for the Closed Scene Composed of Pre-Captured Objects

Fig. L (cont.). More examples of Scene Arrangement on the iG-Synthetic dataset.
Fig. M. More examples of Scene Arrangement on the Fresh-Room dataset.
Fig. N. More examples of Scene Arrangement on the Fresh-Room dataset (extended bedroom scene).
Fig. O. More examples of lighting augmentation on the iG-Synthetic dataset.

Fig. P. More examples of lighting augmentation on the Fresh-Room dataset. Note that for the 3D arrangement visualization, we use the object meshes extracted from our neural implicit renderer.