DORF: Decomposing Object Radiance Field from Supporting Planes

Nelson Nauata
Simon Fraser University
nnauata@sfu.ca

Chen Liu
Meta Reality Labs Research
chenliu91@fb.com

Zhaoyang Lv
Meta Reality Labs Research
zhaoyang@fb.com

(1) Raw image (2) Rendered image (3) Segmentation (4) Object depth (5) Scene editing

Figure 1. Given only localized multi-view images of objects sitting on a planar scene in (1), our method can produce category-agnostic high-quality reconstruction of the object in (2), simultaneously segment the object in 3D in (3), and recover object depth in (4). We can further enable scene editing without the need for any mask inputs, as we copy, resize, and move the Pikachu and put a toy bear extracted from another scene on the reconstructed 3D plane in (5).

Abstract

We propose a novel self-supervised neural reconstruction method that can reconstruct and segment 3D objects at high quality using localized multi-view images. By assuming objects usually sitting on one supporting plane, we represent a 3D scene as one major supporting plane with multiple objects on top, with each entity as a neural radiance field parameterized by a coordinate-based MLP. We jointly reconstruct the 3D objects, the planar structure, and the 3D decomposition via multi-pass differentiable neural radiance field rendering using only images as input. In contrast to the traditional semantic-based 3D object localization and reconstruction methods, our method is completely taxonomy free and can achieve accurate 3D segmentation for a wide variety of opaque objects. We demonstrate that our method can achieve high-quality view-consistent rendering and object segmentation for a wide range of objects in CO3D dataset [18] without using any provided masks. Our method works equally well for objects which can not be detected and localized by semantic-based methods. With accurate geometry, appearance, and decomposition, we can further enable 3D object discovery and scene editing applications.

1. Introduction

With the increasing demand for AR/VR applications, many downstream applications on scene editing and interaction require high quality 3D object segmentation and reconstruction from localized RGB images. A conventional 3D object segmentation and reconstruction pipeline [6, 18, 19] requires pixel-accurate semantic segmentation from a pre-trained instance detection model (e.g. MaskR-CNN [2], PointRend [5]). The reconstruction and segmentation quality of these methods heavily depends on the 2D instance segmentation. Among objects in our surroundings, there are a vast number of "unseen" objects that can not be recognized reliably by the instance detection method, which fails the subsequent reconstruction steps completely.

In contrast, we explore a different geometric representation to discover objects with no dependency on semantic priors. A majority of objects we observe in our daily life lie on a planar structure: a cup on a desk, a desk on the floor, etc. There is a hierarchical geometric structure composed of planar primitives that support various objects on top of it. Compared to the diverse categories of objects varying in shape and appearance, the geometric structure of planes can be more easily and reliably localized using existing methods (e.g. RANSAC, PlaneRCNN [8]). If we can jointly re-

construct the objects with their supporting planar structures reliably, we can decompose them based on their geometric relationship.

In this work, we focus on solving the core module of such geometric representation: self-supervised reconstructing and segmenting multiple objects with their supporting planes. We use neural radiance fields (NeRF) [11] as the differentiable rendering representation. Different from NeRF, we represent a 3d scene as the composition of an instantiated plane and local objects, where each primitive (i.e. a plane or an object) has its own coordinate-based MLP responsible for modeling its unique appearance and geometry. We jointly reconstruct them via volumetric rendering in a mixture model. To successfully decompose individual object primitive from the supporting plane, we propose three main components in the reconstruction process. First, we use a single-layer MLP to parameterize plane geometry to enforce its compactness which drives the decomposition. Second, we utilize a mixture model with multi-pass rendering to jointly optimize the appearance and geometry of all primitives. Third, by assuming that an opaque plane does not occlude the object, we introduce a regularization term to enforce the compactness of the objects that dramatically improves object segmentation.

Our experiments show that we can successfully reconstruct and decompose a variety of objects from their supporting planes. We first validate our method can work at scale on a large variety of objects in Common Objects in 3D dataset [18] and show the effectiveness of our proposed approach with comprehensive ablation studies. We further show our method can tackle small objects on a finite plane in a complex environment. Our method can achieve high quality reconstruction and segmentation of objects when existing object detectors fail to recognize them. Finally, we demonstrate our method can also simultaneously reconstruct multiple objects that share one supporting plane.

To summarize, we provide three contributions:

- We introduce a novel concept to discover 3D objects by jointly reconstructing the objects and their supporting planes from localized images.
- We propose a novel neural reconstruction process with object and plane constraints that can produce high quality 3D segmented object models.
- We demonstrate high quality reconstruction and segmentation of common objects in the CO3D dataset, small objects on a finite plane in a self-captured ego-centric video dataset, and multi-objects on one plane in phone sequences, with superior quality than previous methods including those using semantic inputs.

2. Related Work

Our work is closely related to the recent success of neural rendering techniques that empower high-quality reconstructions from localized images. Among all of them, we are the first work based on neural rendering to achieve general high-quality category agnostic object reconstruction and segmentation with no mask input.

Neural implicit representation: Neural Radiance Fields (NeRF) [11] becomes increasingly popular as image-based reconstruction representation for its reconstruction quality and simplicity to adapt for various imaging inputs. It parameterizes the scene with a coordinate-based MLP. An image is rendered from the integrated radiance field values queried from positions along camera rays. A simple extension to NeRF with additional latent code in input query can cope with images with significant appearance variation and outliers [10] and complex dynamics in 3D videos [7]. Another line of work uses implicit surface models as the scene representation, which can achieve higher reconstruction quality for objects [14, 25] but require ground truth masks as input. Recently several methods unify the strength of radiance field and surface representation in differentiable volumetric rendering, which alleviate the need for accurate object mask as input [15, 21, 24]. All of these approaches share the similarity of using a single coordinate-based MLP as the neural representation for the scene. Though these approaches can represent the scene signals at high quality with positional encoding, they provides no decomposition of the scene: a reconstructed object can not be easily edited or extracted from the scene. Follow-up works represent scene using a mixture of primitives [9] and local MLPs [17] which make it beneficial for real-time rendering, but they produce no reasonable structure to decompose the objects.

Semantic based neural reconstruction: Pre-trained instance detection methods can provide structure decomposition of a scene in the image domain. Several works utilize the category mask to reconstruct and segment objects. Zhang *et al*. [3] uses human-based detector and tracker to localize the dynamic object region and models the scene with dynamic humans separately based on the semantic decomposition. Ost *et al*. [16] build a neural scene graph of reconstructed objects for the 2D localized instances. Reizenstein *et al*. [18] build a dataset of 3D objects using neural rendering for each detected category based on object category mask acquired from [5]. Yang *et al*. [23] demonstrate a composition of object radiance fields with the background model given 2D object masks, which allows further editing. One limitation of such semantic-based methods is that it can only reliable reconstruct objects within certain categories, and will fail for all types of other common objects. In contrast to the existing work using an off-the-shelf instance detection to localize objects, we are category-agnostic and

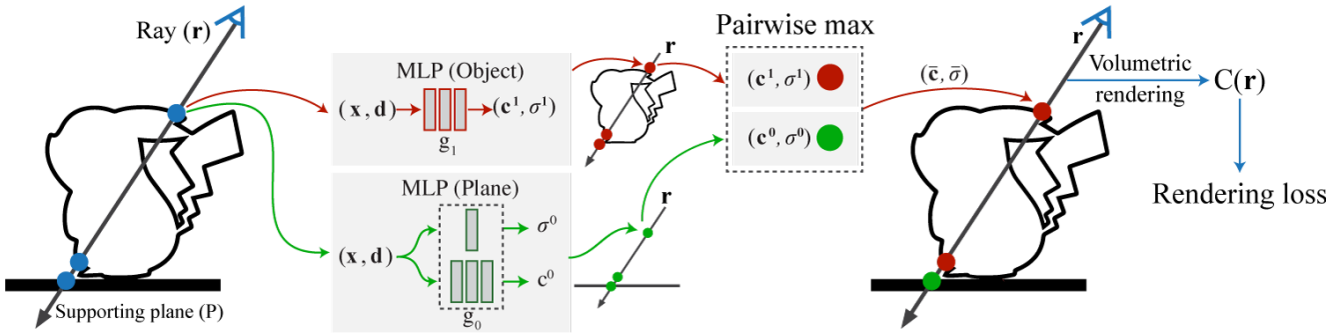


Figure 2. **Method overview.** For a scene with one object and a supporting plane (P), we define the functions g_0 and g_1, represented as MLPs, for mapping spatial locations $\mathbf{x}$ and viewing directions $\mathbf{d}$ to emitted colors and volume densities, $(\boldsymbol{c}^0, \sigma^0)$ for the plane and $(\boldsymbol{c}^1, \sigma^1)$ for the object. The emitted colors and volume densities are combined by a pairwise operation over σ^0 and σ^1, resulting in a new set of colors and densities $(\bar{\boldsymbol{c}}, \bar{\sigma})$, which are used for rendering the color C($\mathbf{r}$) for a ray $\mathbf{r}$.

require no mask input. We can produce category agnostic high-quality 3D reconstruction for objects by jointly reconstructing objects with their supporting plane.

Self-supervised scene decomposition via neural rendering: Several works demonstrated self-supervised rendering-based decomposition of the scene using different photo-geometric inductive biases in the scene. Wu *et al*. [22] exploit symmetry and shading to reconstruct the 3D shape of an object category from raw individual 2D images alone. Yuan *et al*. [27] simultaneously track and reconstruct a single rigid moving object with its background scene. Tschernezki *et al*. [20] use a layered neural rendering representation to segment the moved objects from its background in an ego-centric dynamic video. Guo *et al*. [1] can enable movement of the objects by modeling the light transport as an object-centric neural scattering function in a simulated dataset with control in the light. Niemeyer *et al*. [12, 13] represent the scene as generative compositional radiance fields of objects and background which can be learned using a generative-adversarial training process. Yu *et al*. [26] propose an unsupervised object discovery of object radiance fields based on a compositional rendering of NeRF. These methods [12, 13, 26] all require pre-training on a large amount of data with different configurations of object, camera and background to exploit object feature and attention, which has not demonstrated real-world object discovery performance. In contrast, our method does not require pre-training. Within these existing category-agnostic object decomposition work, we are the first to demonstrate our method can successfully handle a vast amount of objects using real-world videos.

3. Method

Our goal is to infer a decomposition of the 3D scene as a hierarchy of primitives from image observations. We assume that a 3D scene is composed of a supporting plane P and K foreground objects $\{O_k\}_{k=1}^K$. We can represent complex scenes hierarchically by further decomposing foreground objects. Sec. 3.1 explains our representation in details. In Sec. 3.2, we discuss how to jointly render the supporting plane and foreground objects using a mixture volumetric rendering model and train them using image reconstruction loss. To increase the decomposition quality, we add an object-centric regularization term in Sec. 3.3.

3.1. Representation

We use neural radiance fields (NeRF) [11] to represent the supporting plane and all objects. For each NeRF, an MLP network parameterizes a continuous mapping $g : (\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma)$ from spatial location $\mathbf{x} \in \mathbf{R}^3$ and viewing direction $\mathbf{d} \in \mathbf{R}^2$ to emitted color $\mathbf{c} \in \mathbf{R}^3$ and volume density $\sigma \in \mathbf{R}$. Specifically, we represent the supporting plane P by g_0 and foreground objects $\{O_k\}_{k=1}^K$ by $\{g_k\}_{k=1}^K$.

Plane Parameterization: To ensure the supporting plane has a compact planar geometry, which is crucial for successful decomposition as shown in Sec. 4.1, we constrain the volume density of g_0 such that it has only one learnable linear layer and a *logistic density distribution* function as its activation [21]. In theory, an infinite plane has a linear Signed Distance Field $f(\mathbf{x}) = \mathbf{n} \cdot \mathbf{x} - d$ where $\mathbf{n} \in \mathbf{R}^2$ is the plane surface normal and $d \in \mathbf{R}$ is the plane offset. A linear MLP layer can perfectly model $f(\mathbf{x})$. The *logistic density distribution* activation function converts $f(\mathbf{x})$ to volume density $\sigma(\mathbf{x})$ by,

$$\sigma(\mathbf{x}) = \frac{s e^{-s f(\mathbf{x})}}{(1 + e^{-s f(\mathbf{x})})^2} \tag{1}$$

where learnable parameter s controls the sharpness of the plane geometry. The volume density $\sigma(\mathbf{x})$ is the largest when $\mathbf{x}$ lies on the plane (i.e., $f(\mathbf{x}) = 0$) and it approaches to 0 when $\mathbf{x}$ moves away from the plane.

3.2. Scene Mixture Rendering

We propose a mixture rendering model to compose the supporting plane and objects into one holistic neural radiance field, and jointly optimize their geometries and appearances using image observations. Given the supporting plane P and K objects $\{O_k\}_{k=1}^K$, our scene mixture model aggregates them into one global NeRF by taking the maximum volume density and corresponding emitted color at each spatial location. Specifically, for spatial location $\mathbf{x}$, NeRF mapping functions $\{g_k\}_{k=0}^K$ predict density values $\{\sigma^k\}_{k=0}^K$, and the maximum density has index $\bar{k} = argmax_k(\{\sigma^k\}_{k=0}^K)$. $\bar{k}$ is the same for all viewing directions at the same $\mathbf{x}$. We then use the $\bar{k}$th NeRF to estimate the aggregated emitted color $\bar{\mathbf{c}}$ and volume density $\bar{\sigma}$ at spatial location $\mathbf{x}$ and viewing direction $\mathbf{d}$.

We sample M positions along a camera ray discretely as $\mathbf{r} = \{\mathbf{o}+t_i\mathbf{d}\}_{i=1}^M$, where $\mathbf{o}$ is the camera center and t_i is the depth along the ray. We compute the expected color $C(\mathbf{r})$ at camera ray $\mathbf{r}$ using the volumetric rendering equation,

$$C(\mathbf{r}) = \sum_{i=1}^{M} T_i(1 - \exp(-\bar{\sigma}_i\delta_i))\bar{\mathbf{c}}_i \tag{2}$$

where $T_i = \exp(-\sum_{j=1}^{i-1} \bar{\sigma}_j\delta_j)$ and δ_i is the distance between two adjacent samples along a ray.

We train our model by comparing N rendered images $I_{i=1}^N$ to ground truth images $\hat{I}_{i=1}^N$ using the image reconstruction loss $\mathcal{L}_{recon} = \sum_{i=1}^N ||I - \hat{I}||^2$. We do not require 3D geometry or segmentation supervision. We can recover 3D instance segmentation when we find the primitive with the maximum density at each spatial location $\mathbf{x}$.

3.3. Object-centric Regularization

For high-quality segmentation, we need to enforce proper constraints to avoid the ambiguities from $K + 1$ NeRF primitives at the same region. In an ideal scene decomposition, decomposed primitives should have compact geometries and appearances. The one-layer MLP formulation introduced in Sec. 3.1 models plane geometry with the optimal compactness, which is the driving factor for a successful decomposition. For objects, we add a regularization term to improve the compactness of object geometries.

The regularization is based on a justified assumption that a camera ray does not intersect with any objects if it hits the supporting plane first. For camera ray $\mathbf{r} = \{\mathbf{o} + t_i\mathbf{d}\}_{i=1}^M$, we determine its first intersection point by finding the peak of the emission-absorption product such that $\bar{i} = argmax_i(T_i(1 - \exp(-\bar{\sigma}_i\delta_i)))$. We collect all the camera rays that hit the supporting plane first as $\mathcal{R}^P$, and we regularize the object MLPs by minimizing their weights,

$$\mathcal{L}_{reg} = \sum_{\mathbf{r}\in\mathcal{R}^P} \sum_{k=1}^{K} \sum_{i=1}^{M} \sigma_k \tag{3}$$

(a) $w_{reg} = 0$ (b) $w_{reg} = 0.1$ (c) $w_{reg} = 1.0$

Figure 3. **The effect of object-centric regularization.** As we gradually increase the regularization weight during training, we get sharper object masks.

This regularization is effective to remove cloudy object points and improve object segmentation boundary as shown in Fig. 3.3. Our final training loss is $L = \mathcal{L}_{recon} + w_{reg}\mathcal{L}_{reg}$, where w_{reg} is a scalar weight for $\mathcal{L}_{reg}$.

3.4. Implementation Details

Initialization: Instead of training all $K + 1$ NeRFs from scratch, which is prone to stay in local minima, we deploy a stratified training strategy. We first only train one global NeRF to estimate a rough geometry for the entire scene. After enough iterations, we decompose the rough geometry (i.e. point cloud) from NeRF into a set of primitives to initialize $K + 1$ NeRFs. To be specific, we extract the point cloud from estimated density volume and the supporting plane using RANSAC. For the rest of the point cloud not belonging to a plane, we deploy K-means to estimate the number of objects for initialization. For each object cluster, we compute a bounding box by finding extremes with 80% margin along the plane normal direction and two plane tangent directions. We run K-means multiple times increasing K from one to ten till object box is smaller than a size threshold. Please refer to the supplementary material for more details. After initialization, we train all $K + 1$ NeRFs jointly using the method described in Sec. 3.2 and 3.3.

Finite plane: In Sec. 3.1, we assume the supporting plane is infinite (i.e., it occupies the entire image screen) so that we can use one-layer MLP to model its 3D geometry. To handle finite planes in real-world images, we initialize a foreground box that tightly encloses all the object boxes for the trained global NeRF volume at the initialization stage, and ignores points outside the box in further training. This simple yet effective strategy filters out background regions and allows the network to focus on the object decomposition in a local region. Although the foreground box often does not cover the entire plane, the enclosed region is enough for successful object discovery as shown in Sec. 4.2.

Hierarchical volume sampling: We use the same coarse-to-fine volume sampling strategy for mixture rendering as NeRF [5], except in the fine sampling step. We predict the importance sampling based on the coarse sample weights for each individual object volume, and aggregate all the sampled points from all primitives in the mixture rendering.

Network architecture: We use the same network architecture as NeRF for each object primitive and plane, except that the density prediction is a single linear layer with a *logistic density distribution* function (eq.1) as its activation for plane primitive. We use shared weights for coarse and fine MLPs in each primitive.

Training Details: We trained the global NeRF in initialization stage with $200K$ iterations and then we train our decomposed volumes jointly with another $200K$ iterations. We use ADAM optimization [4] with $\gamma = 0.1$, and learning rate $1e^{-3}$ for supporting plane density MLP and learning rate $5e^{-4}$ for the rest parameters. Object regularization weight w_{reg} starts with $1e^{-4}$ and double every $2K$ iterations till it reaches 1. Training time is around 15 hours per scene running one a single GPU, inference time is around 1 hour per scene on the entire test views. During training, we sample 512 rays in each image view and 128 "coarse" points and 128 "fine" points along each ray. We initialize NeRF weights from scratch and $s = 1.0$.

4. Experiments

We experimented on a variety of real-world data to evaluate the performance of our method and validate our system designs. First, we evaluate our method on a large amount of objects using Common Objects in 3D (CO3D) dataset [18] in Sec. 4.1. In Sec. 4.2, we analyze our DORF performance handling small objects and a finite plane in a complex environment observed from an ego-centric device. Finally, we demonstrate that our method can discover multiple objects sharing a single supporting plane in Sec. 4.3. Please check our supplementary materials for qualitative results of our method on more data and reconstruction analysis in 3D.

4.1. Evaluation on Common Objects at Scale

We utilize the CO3D dataset [18] to demonstrate our method's capability to handle a large variety of objects at scale. This dataset also provides pseudo ground truth masks to evaluate the segmentation quality. We provide a comprehensive ablation study of our method and quantitative evaluation of our method compared to baselines.

Dataset: We use the CO3D test set for the single object evaluation. We do not use their training set since our method does not require additional training. We report the same metrics for comparing with competing methods on object reconstruction (Tab. 1). We select two subsets from the test set for more precise ablation studies:

- **CO3D subset with precise masks:** We found not all masks used as pseudo ground truth acquired from PointRend are precise enough. Our segmentation mask output can be of higher quality in many sequences, as indicated in Fig. 5. We use this subset for ablation studies on segmentation quality in Tab. 1. The *intersection over union* (IoU) measures segmentation mask quality.
- **CO3D subset with challenging initialized planes:** This subset was selected based on the estimated plane parameter using RANSAC, we select those with "inaccurate" plane parameters to evaluate our plane parameter optimization strategy.

Both subsets contain 10 sequences from 5 different categories. We use "apple", "bowl", "mouse", "skateboard", "suitcase"for the first subset and "ball", "donut", "hydrant", "orange", "remote" for the second. All models are trained on 80 views of a sequence and the metrics are computed on the 20 held out test views.

Metrics: We evaluate the reconstructed object quality using photometric metrics PSNR and LPIPS, and the segmentation quality using IoU. For all object-only evaluations in Tab. 1 and Tab. 2, we compare the rendered object to the corresponding ground truth object while ignoring background pixels.

Ablation Study: To validate the contributions of each proposed module, we perform the following ablation studies.

- **Initial plane extraction using RANSAC (INIT):** This is our initialization step describe in Sec.3.4. We extract the rough geometry from a trained model using global NeRF, and estimate the plane parameters using RANSAC.
- **Mixture Renderer (MR):** We parameterize the object

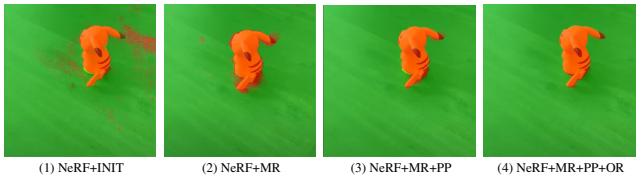


Figure 4. **A visual comparison of the object segmentation quality for ablation study.** We visualize the object with its mask in Red and the plane in Green. DORF (NeRF+MR+PP+OR) produces accurate segmentation with a sharp boundary.

Table 1. **Ablation study of proposed modules on the CO3D subset with precise masks.** We evaluate our proposed modules using NeRF with primitive initialization (INIT), and progressively adding Mixture Renderer (MR), Plane Parametrization (PP) and Object Regularization (OR). We report all metrics for the object region. Our final model DORF (NeRF+MR+PP+OR) significantly outperforms the NeRF baseline.

Model	PSNR ↑	LPIPS ↓	IoU ↑
NeRF+INIT	10.1	0.52	0.24
NeRF+MR (ours)	9.9	0.53	0.24
NeRF+MR+PP (ours)	14.9	0.46	0.45
NeRF+MR+PP+OR (ours)	**21.5**	**0.23**	**0.82**

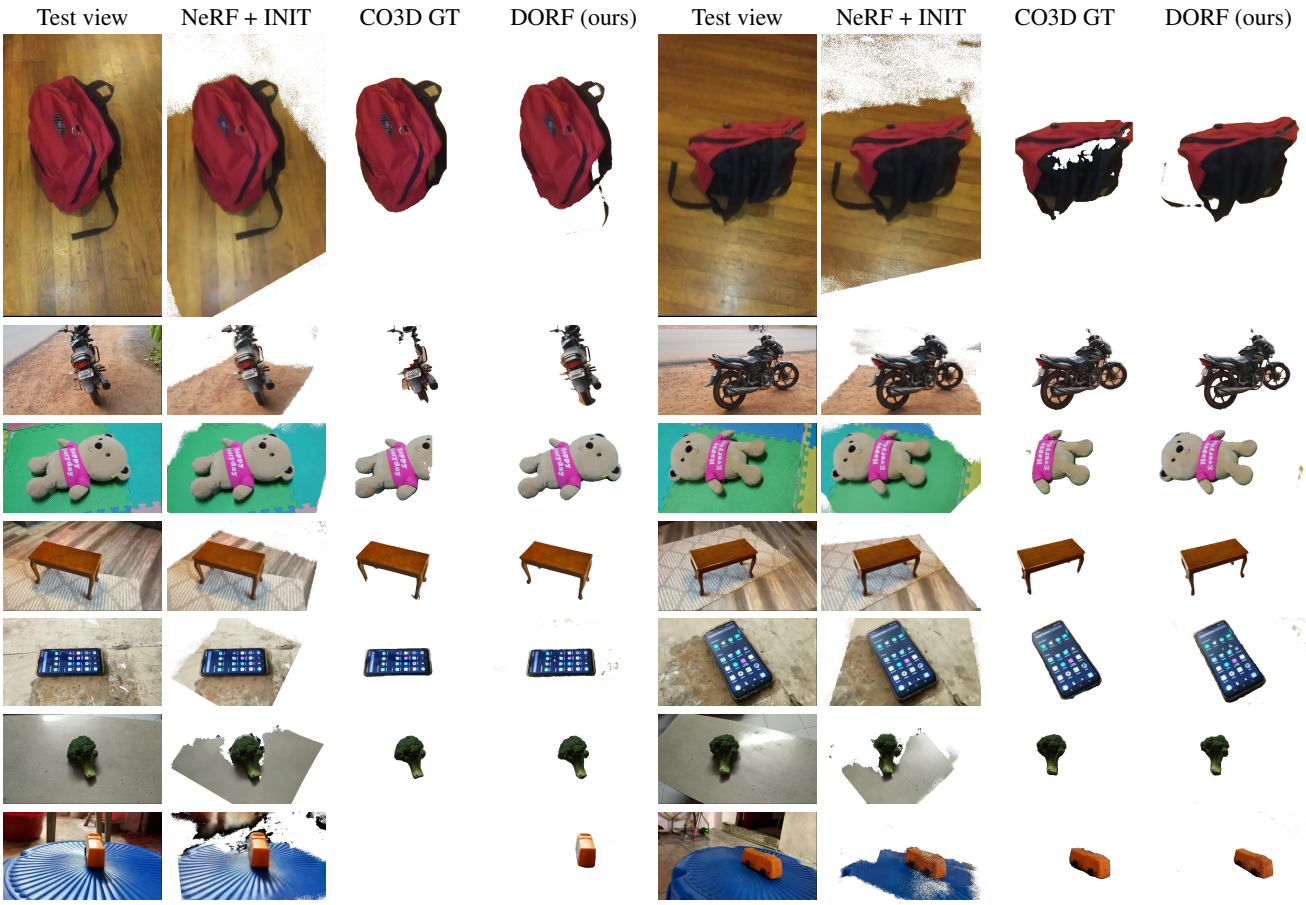


Figure 5. **Qualitative results of our reconstruction on different objects in CO3D [18].** Our method can handle objects with various shape and appearance, and produce high quality reconstruction of the objects even with thin structures. The object masks by NeRF + INIT are often noisy and contain most of the background portion. Compared to the pesudo ground truth provided in CO3D [18] (CO3D GT) which is acquired using instance segmentation, our reconstruction can potentially produce more accurate and consistent object segmentation across different views. The segmentation image is white if there is no predicted object mask.

and plane using two separate MLPs and jointly optimize them with the mixture rendering as proposed in Sec. 3.2.

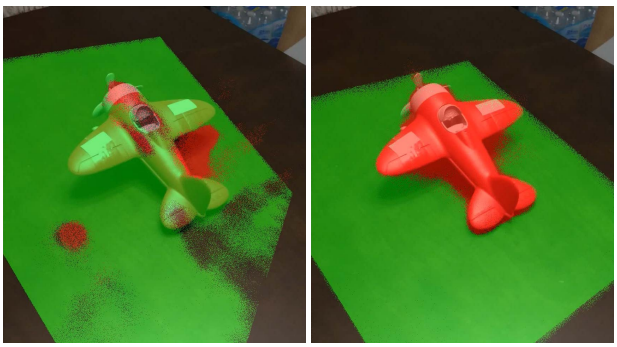
- **Plane Parameterization (PP):** We paramterize the plane using a compact planar parameterization in Sec. 3.1.
- **Object Regularization (OR):** We add the regularization loss (eq. 3) in Sec. 3.3.

Tab. 1 shows the quantitative ablation study using the CO3D subset with precise masks. We use a Pickachu example to demonstrate the ablation qualitatively in Fig. 4. Overall, NeRF with initialized plane geometry (INIT) can not provide good reconstruction and segmentation of the object. From Fig. 4, we can see mixture rendering (MR) can reduce the noisy segmentation of the objects in the planar region, compared to extracting planes using RANSAC from a geometry model produced by NeRF (INIT). However, MR alone is still not sufficient to improve the quality overall, as indicated in Tab. 1. Our Plane Parameterization (PP) improves the object mask with much less noise. The Object Regularization (OR) further improves reconstruction and segmentation of objects as shown in Fig. 4 and Tab. 1.

Table 2. **Ablation study of plane parameter optimization on CO3D subset with challenging initialized planes.** We report all metrics for object region. Optimizing the plane can improve the segmentation quality of challenging cases by a large margin.

Model	PSNR ↑	IoU ↑
w/o plane optimization	18.1	0.30
w/ plane optimization	20.5	0.42

The effect of plane optimization: We observe that primitive initialization can provide good plane initialization for most scenes, so that DORF is able to achieve successful decomposition results without further optimizing the plane geometry. For scenes where the plane is not successfully



(1) NeRF+INIT (2) DORF (ours)

Figure 6. **A segmentation visualization of the plane optimization.** We visualize the reconstructed object mask in Red and the local planar region in Green. For a plane with noisy initialization (1), DORF can estimate a more accurate supporting plane by jointly optimizing the plane parameters in the reconstruction (2).

initialized, jointly optimizing plane parameters encoded by plane density MLP can be helpful. We perform the ablation study of plane optimization using CO3D subset with challenging plane initialization. In Tab. 2, we show that our plane parameter optimization leads to better overall scores. As shown in Fig. 6, our optimization strategy helps recover from cases where RANSAC fails to estimate accurate plane parameters. However, we noticed that for cases when the plane parameters are severely inaccurate our method struggles to recover the correct plane parameters. For those cases, we obtain "empty" object masks, which lead to a low average score as shown in Table 2.

Results compared to semantic-based reconstruction: In Fig. 5, we demonstrate a few examples of the segmentation mask produced by DORF compared to the instance segmentation mask acquired from PointRender [5], which is used as pseudo ground truth for all mask evaluation in CO3D dataset. DORF can potentially produce more accurate and consistent instance mask with varying views.

4.2. Evaluation of Objects on a Finite Plane

Most of the videos in CO3D are object-centric videos, in which almost all objects and planes are salient in views and not representative for objects observed from ego-centric AR devices. To evaluate how our method works for ego-centric videos on a local finite plane, we captured a few wide-angle egocentric videos observing the objects on a desk.

Dataset: We capture five sequences of different objects (a dinosaur toy, a clock, a birdhouse, a bowl, and a mug) on a desk from an ego-centric AR glass. We record the entire video in a room with a motion capture system. In each sequence, we observe the object with a full circle around the desk. For the entire capture, we have the 3D ground truth for object pose, the full camera trajectory, and the object's 3D digital model. We acquire the ground truth object mask by projecting the object model into each image view. Fig. 7 shows a few examples of the object mask ground truth. For each sequence, we split the trajectory of training and testing views with a ratio of 5:1. We provide more details about this dataset in supplementary materials.

Table 3. **Quantitative evaluation of DORF for objects on a finite plane.** We compute the average statistics for all test views of the five object sequences.

Model	PSNR ↑	LPIPS ↓	IoU ↑
NeRF + INIT	12.69	0.56	0.11
NeRF + MaskRCNN	25.53	0.12	0.46
DORF (ours)	**28.19**	**0.06**	**0.72**

Baselines: We compare to the following baselines:

- **NeRF + INIT:** Same as the ablation baseline in Sec. 4.1, we estimate the planar mesh region using RANSAC from the geometry extracted by NeRF. This is the primitive initialization stage of our approach and serves as a common baseline to acquire objects on a plane using NeRF.
- **NeRF + MaskRCNN [2]:** We use estimtated object mask from pretrained Mask-RCNN for object segmentation. Similar as the NeRF object baseline in Co3D [18], this serves a standard baseline for 3D object segmentation when its semantic information is available.

Metrics: We use the evaluation metric as CO3D ablation in Tab. 1. We evaluate object reconstruction quality using PSNR and LPIPS, and the segmentation quality in IoU.

Results: We demonstrate the comparisons quantitatively in Tab. 3 and qualitatively in Fig. 7. For all sequences, DORF can successfully localize the desk plane in the complex environment, reconstruct and segment the object in 3D. Our quantitative evaluation in Tab. 3 shows DORF can produce better reconstruction and segmentation for objects which can not be trivially extracted from a global NeRF model or the state-of-the-art pre-trained instance segmentation method. DORF can work successfully even for small objects (e.g. clock and toy dinosaur). For rare object category (e.g. birdhouse and toy dinosaur), DORF can work equally well when MaskRCNN is struggling to localize and segment the correct instance across different views.

4.3. Multiple Objects on a Plane

We further demonstrate our method can tackle multiple objects on a plane in Fig. 8. The sequence contain 200 training frames. The objects are supported by one infinite plane. Some objects may be partially excluded from the camera view for some frames or occluded by another object. The

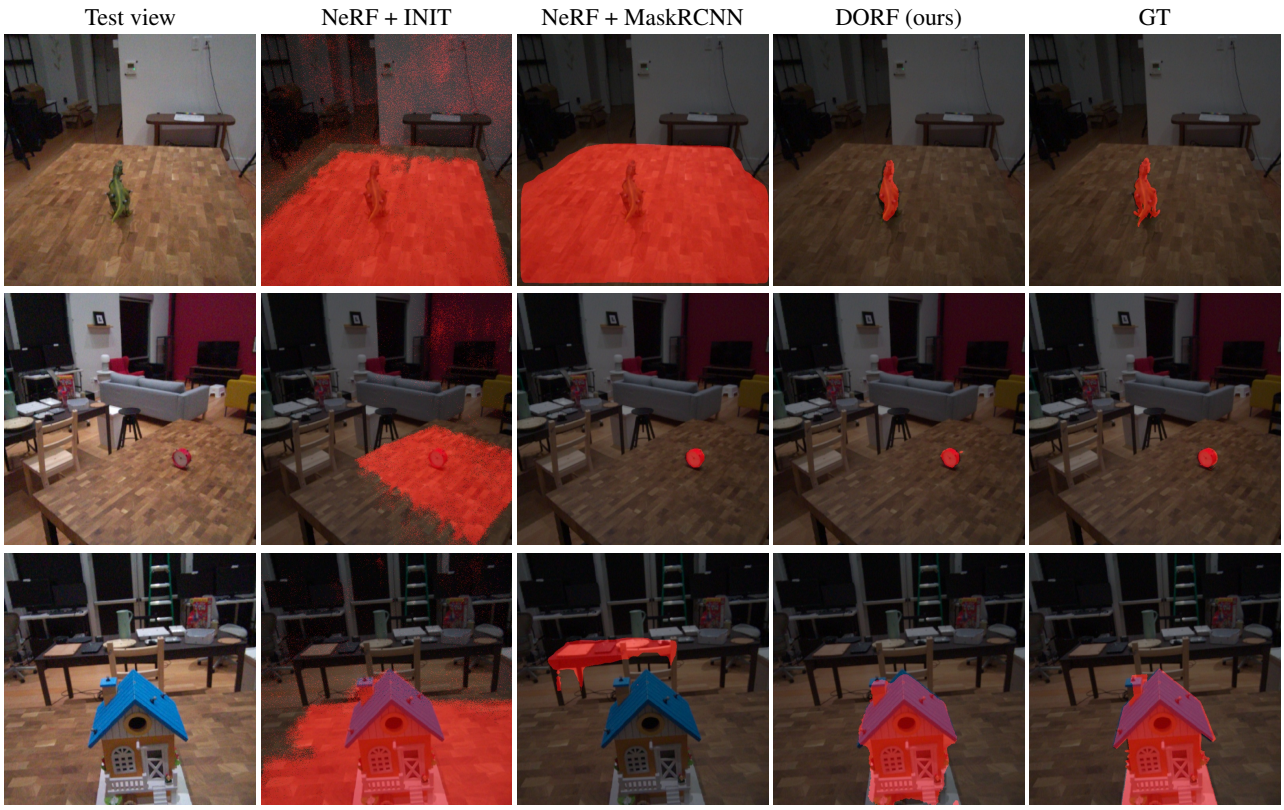


Figure 7. **Qualitative comparison of object segmentation on a finite plane.** The mask predictions from all methods are visualized in Red. Compared to baselines, DORF can produce significantly better reconstruction and segmentation, even for very small objects. DORF can also produce high-quality masks for objects when semantic-based approaches failed to detect them.

Figure 8. **A visualization of DORF handling multiple objects on the plane.** We visualize the different reconstructed objects with masks in different colors and the planar region in Green.

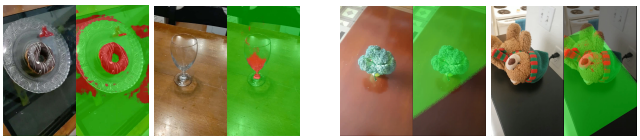


Figure 9. **Example of failure cases.** We visualize the reconstructed object masks in Red and the planar region in Green.

book is also a thin structure on the plane. Despite of the challenges, our method successfully decompose all objects.

4.4. Limitations and Future Work

Our method cannot tackle transparent objects or support planes. When the NeRF reconstruction is poor or the supporting plane is very small, our initialization of plane can be off significantly and break the whole optimization process. Fig. 9 shows a few failure cases.

This paper is only an initial attempt towards 3D object discovery in the wild. We have not explored more complex scenarios that contain cluttered objects with different levels of supporting planes. We believe that our work is an essential step towards this goal and leave this for future works.

5. Conclusion

We have proposed a novel concept to simultaneously reconstruct and segment objects from their supporting planes using multi-view images. To realize this concept, we use a neural rendering based approach to jointly reconstruct the objects together with their supporting planes. We carefully analyze the effectiveness of our approach on a variety of real-world datasets. Our results demonstrate the potential of our approach to achieve 3D object discovery without the dependency on semantic priors. We believe that our findings can further inspire future works to build novel systems for 3D object discovery in the wild.

References

[1] Michelle Guo, Alireza Fathi, Jiajun Wu, and Thomas Funkhouser. Object-centric neural scene rendering. *arXiv preprint arXiv:2012.08503*. 3

[2] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *ICCV*, 2017. 1, 7

[3] Zhang Jiakai, Liu Xinhang, Ye Xinyi, Zhao Fuqiang, Zhang Yanshun, Wu Minye, Zhang Yingliang, Xu Lan, and Yu Jingyi. Editable free-viewpoint video using a layered neural representation. In *ACM SIGGRAPH*, 2021. 2

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[5] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross B. Girshick. Pointrend: Image segmentation as rendering. In *CVPR*, 2020. 1, 2, 4, 7

[6] Kejie Li, Daniel DeTone, Yu Fan Steven Chen, Minh Vo, Ian Reid, Hamid Rezatofighi, Chris Sweeney, Julian Straub, and Richard Newcombe. Odam: Object detection, association, and mapping using posed rgb video. In *ICCV*, pages 5998–6008, 2021. 1

[7] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, and Zhaoyang Lv. Neural 3d video synthesis. *arXiv preprint arXiv:2103.02597*, 2021. 2

[8] Chen Liu, Kihwan Kim, Jinwei Gu, Yasutaka Furukawa, and Jan Kautz. Planercnn: 3d plane detection and reconstruction from a single image. In *CVPR*. 1

[9] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering. In *ACM SIGGRAPH*, 2021. 2

[10] Ricardo Martin-Brualla, Noha Radwan, Mehdi Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the wild: Neural radiance fields for unconstrained photo collections. In *CVPR*, 2022. 2

[11] 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. In *ECCV*, 2020. 2, 3

[12] Michael Niemeyer and Andreas Geiger. Campari: Camera-aware decomposed generative neural radiance fields, 2021. 3

[13] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *CVPR*, 2021. 3

[14] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *CVPR*, June 2020. 2

[15] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *ICCV*, 2021. 2

[16] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. In *CVPR*, 2021. 2

[17] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *ICCV*, 2021. 2

[18] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *ICCV*, 2021. 1, 2, 5, 6, 7

[19] Martin Runz, Kejie Li, Meng Tang, Lingni Ma, Chen Kong, Tanner Schmidt, Ian Reid, Lourdes Agapito, Julian Straub, Steven Lovegrove, and Richard Newcombe. Frodo: From detections to 3d objects. In *CVPR*, June 2020. 1

[20] Vadim Tschernezki, Diane Larlus, and Andrea Vedaldi. NeuralDiff: Segmenting 3D objects that move in egocentric videos. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2021. 3

[21] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *Advances in Neural Information Processing Systems*, 2021. 2, 3

[22] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3D objects from images in the wild. In *CVPR*, 2020. 3

[23] 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 *ICCV*, October 2021. 2

[24] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *arXiv preprint arXiv:2106.12052*, 2021. 2

[25] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33, 2020. 2

[26] Hong-Xing Yu, Leonidas J. Guibas, and Jiajun Wu. Unsupervised discovery of object radiance fields. *arXiv preprint arXiv:2107.07905*, 2021. 3

[27] Wentao Yuan, Zhaoyang Lv, Tanner Schmidt, and Steven Lovegrove. Star: Self-supervised tracking and reconstruction of rigid objects in motion with neural rendering. In *CVPR*, pages 13144–13152, June 2021. 3