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

A. Overview

In this document, we provide technical details in support of our main paper. Below is a summary of the contents.

Sec. B: Description of supplementary video;
Sec. C: Additional applications enabled by our method;
Sec. D: Mesh reconstruction by our method and baselines;
Sec. E: Additional details on the primitive initialization;
Sec. F: Additional details on the code and data release;
Sec. G: Additional results on CO3D dataset.

B. Video

We encourage the reader to watch our supplementary video, where we visualize the following results:

- A comparison of DORF (Ours) against traditional semantic-based 3D object localization methods i.e. Mask-RCNN [1]. We show that our method is capable of achieving high-quality 3d aware object decomposition of unknown categories, which are taxonomy-free and consistent across frames. Meanwhile, Mask-RCNN struggles to recognize the object and shows artifacts such as heavy flickering, fuzzy object boundary and missing parts.

- Additional visualization of our results on CO3D [5], showing that our method can work at scale and in the wild for samples from the CO3D dataset. We compare our method against PointRend [2], which is currently the state-of-the-art for semantic-based 2D segmentation.

- Full sequence for a scene containing multiple objects. Our method can also handle scenes with multiple objects. We show in the video a sample scene containing three objects captured with an iPhone camera.

- Examples of applications that our method can enable without relying on any segmentation masks as input. More precisely, we show one demo for object removal, where our method can accurately remove a target object from a scene, and another demo for editable scene rendering, where we manipulate the object in the scene by shrinking, moving, and rotating it.

(a) Object removal (b) Editable scene rendering

Figure 1. **Examples of application enabled.** DORF can be used for removing an object from the scene (left) and for manipulating the decomposed object from its supporting plane (right).

C. Additional Applications

We demonstrate the potential impact of our method and the applications it can enable without relying on any input segmentation masks. The applications we present are the following.

Object removal: DORF can be directly used for removing objects from a target scene without any input masks. We can simply decompose the object from the supporting plane using DORF and directly utilize the primitives MLPs for rendering the individual primitives. In Figure 1a, we use the decomposed MLPs for rendering the plane plus, the object on the left half of the image and only the plane on the right half. We highlight that our method does not tackle occluded regions nor relighting, which can cause artifacts on the floor when the object is removed or edited. We leave this as future work. The full sequence is available in the supplementary video.

Editable scene rendering: The trained MLPs obtained from DORF can be easily used for manipulating the scene (e.g. inserting, shrinking, moving, and rotating objects.)

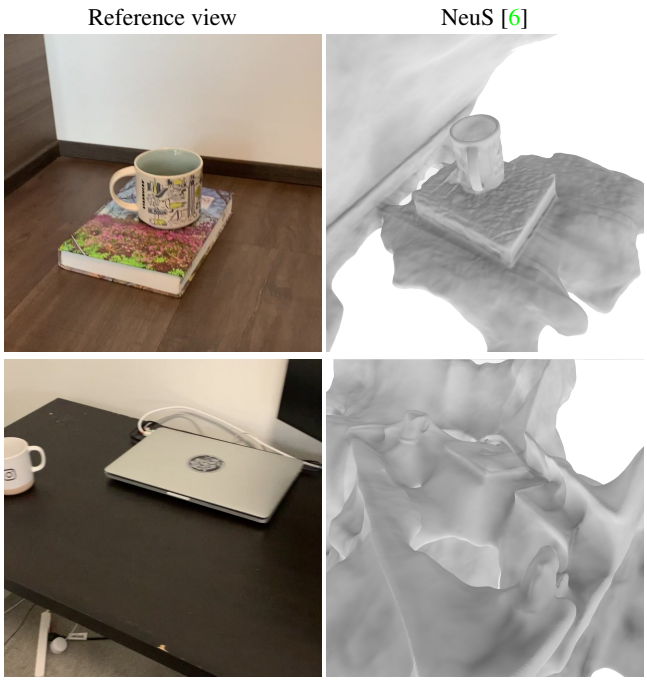


Figure 2. **Corrupt planar surface from NeuS [6].** Examples of corrupt surfaces produced by NeuS when reconstructing common textureless surfaces in indoor scenes. The scenes were captured using an iPhone camera.

and rendering the object in different views. In Figure 1b, we insert another object in the scene rendered from a different view and place it in a different position on top of the estimated supporting plane by DORF. The full sequence is available in the supplementary video.

D. Alternative Scene Representation

In the main paper, we introduce a new concept to discover 3D objects by jointly reconstructing the objects and their supporting planes from localized images using a backbone model. The motivation behind adopting NeRF [3] as our main backbone is due to its robustness to reconstruct a rough geometry for textureless surfaces, which are commonly present in the supporting planes in many of the indoor scenes (e.g. tabletops with small objects on top). Surface reconstruction methods (e.g. NeuS [6], UNISURF [4]) are capable of reconstructing impressively smooth meshes in controlled scenes however, it surfers when reconstructing textureless surfaces leading to corrupt planar surfaces (see Fig. 2), which makes it difficult to extract the major supporting planes from a scene. Though, we foresee our method subsuming the signed distance function (SDF) representation from NeuS [6] and potentially achieving higher-quality reconstruction mesh in controlled scenes. To show its potential, we will present next an additional ablation study focusing on NeuS [6], while leaving for future work its full integration into our system.

Additional baselines:

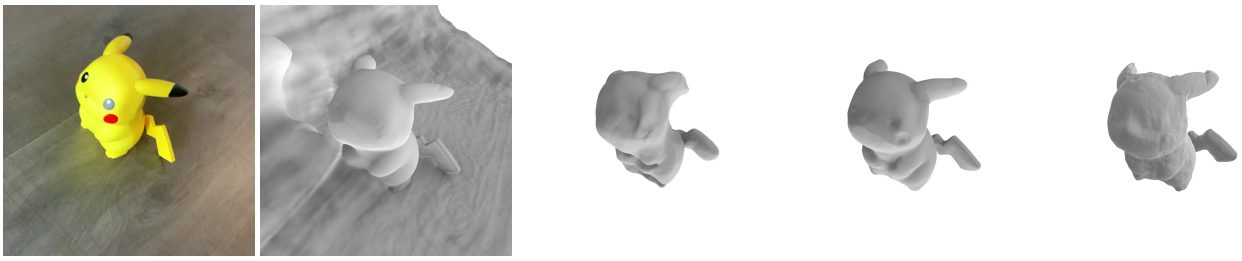
- **NeuS [6]:** Proposed by Wang *et al*., NeuS [6] is one of the state-of-the-arts methods for performing surface reconstruction, in our experiments, we utilize the official code release provided by the authors for replicating it.
- **NeuS [6] + MaskRCNN [1]:** We utilize the same official code release by Wang *et al*. and provide Mask-RCNN [1] segmentation masks for the object as the input masks when training NeuS.
- **NeuS [6] + DORF (Ours):** The same as NeuS + Mask-RCNN, except that we utilize the masks obtained by our method (DORF) as input object masks for training NeuS.

Analysis on the object reconstruction quality:

- As shown in Figure 3b, NeuS [6] alone without any input masks, is susceptible to textureless surfaces (e.g. "Pikachu scene"), which makes it non straightforward to decompose the geometry from its supporting plane. Even using Mask-RCNN for decomposing the object, the reconstructed mesh, Figure 3c, is lower quality, due to missing parts and heavy flickering (inconsistencies) in the segmentation masks from Mask-RCNN.
- Using DORF, Figure 3e, we can decompose the object mesh from its supporting plane successfully. Plus, using NeuS + DORF, Figure 3d, we obtain higher-quality mesh plus object decomposition, which indicates the potential of NeuS as a backbone candidate for our system.
- We noticed that NeuS is more sensitive to noisy segmentation masks, which hurts its reconstruction quality and increase its convergence time. Noisy segmentation masks can happen specially when dealing with complex scenes containing small objects. For example, NeuS does not converge when using inaccurate Mask-RCNN masks as input as shown in Figure 4c. Meanwhile, our method still generates a reasonable reconstruction mesh since we use NeRF as our backbone (Figure 4e.).

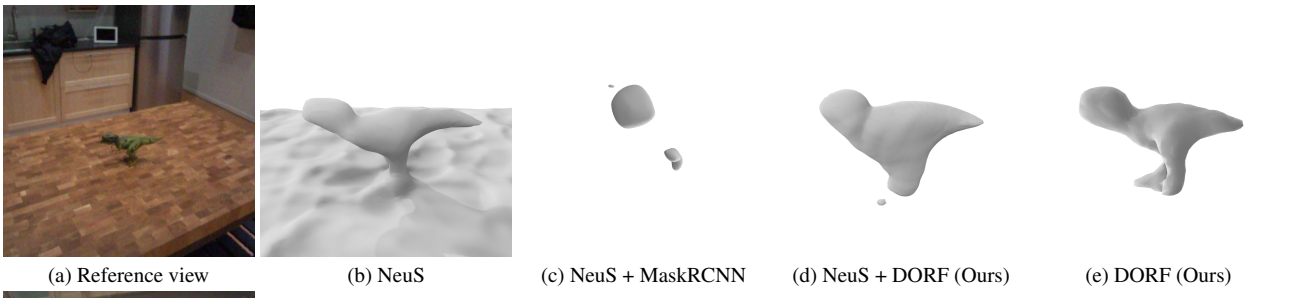
E. Primitives Initialization Details

After the warm-up period where we train one NeRF for the entire scene, we initialize $K + 1$ NeRF models for the supporting plane and K objects. We extract 10,000 points by finding the sample point with the largest density along each camera ray. We fit the supporting plane using RANSAC with 100 iterations and 0.02 as distance threshold (camera centers are normalized onto a unit sphere). We



(a) Reference view (b) NeuS (c) NeuS + MaskRCNN (d) NeuS + DORF (Ours) (e) DORF (Ours)

Figure 3. **Reconstruction results for an object on an infinite plane.** We show that NeuS alone struggles to reconstruct the major supporting plane (a), while DORF is capable of decomposing the object from the major supporting plane (e). When using the mask obtained by DORF, NeuS + DORF (d) successfully achieves decomposition and high-quality mesh reconstruction, while NeuS + Mask-RCNN (c) generates a low-quality mesh with missing parts and fuzzy boundaries.



(a) Reference view (b) NeuS (c) NeuS + MaskRCNN (d) NeuS + DORF (Ours) (e) DORF (Ours)

(f) Reference view (g) NeuS (h) NeuS + MaskRCNN (i) NeuS + DORF (Ours) (j) DORF (Ours)

Figure 4. **Reconstruction results for an object on a local finite plane.** We show that NeuS + Mask-RCNN struggles to converge (c) and (h) for objects from unknown categories (i.e. dinosaur toy and birdhouse.). NeuS + DORF is sensitive to the input mask quality, which is lower in a complex scene with small objects. Meanwhile our method seems more robust as it utilizes NeRF [3] as backbone.

remove the inlier points of the supporting plane and run K-Means on the remaining points to compute object boxes. 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 increase K from one to ten till the smallest object box contains less than 2,000 points.

F. Code and Data

We obtained the necessary consents where we collected the data. For thirdparty datasets from CO3D [5], we refer to them for details. Our dataset does not contain any human identified information or offensive content.

For our new dataset used in Section 4.2, we use Optitrack to obtain accurate 3D poses for objects and the egocentric camera. Each object is pre-scanned and has a high-quality ground truth mesh model as its digital twin. We can project the mesh model to get precise object masks. Due to the proprietary of the high quality 3D digital model, we cannot promise to release the full dataset at the time of submission. We believe such 3D ground truth for real-world data will be of value for academic community. We are working with the legal on the possibility to release this ground truth data for research purpose.

We intend to release the code as well, upon the condition of being granted permission from the legal.

G. Additional Results

In Figs. 5-8, we provide additional results on the CO3D dataset.

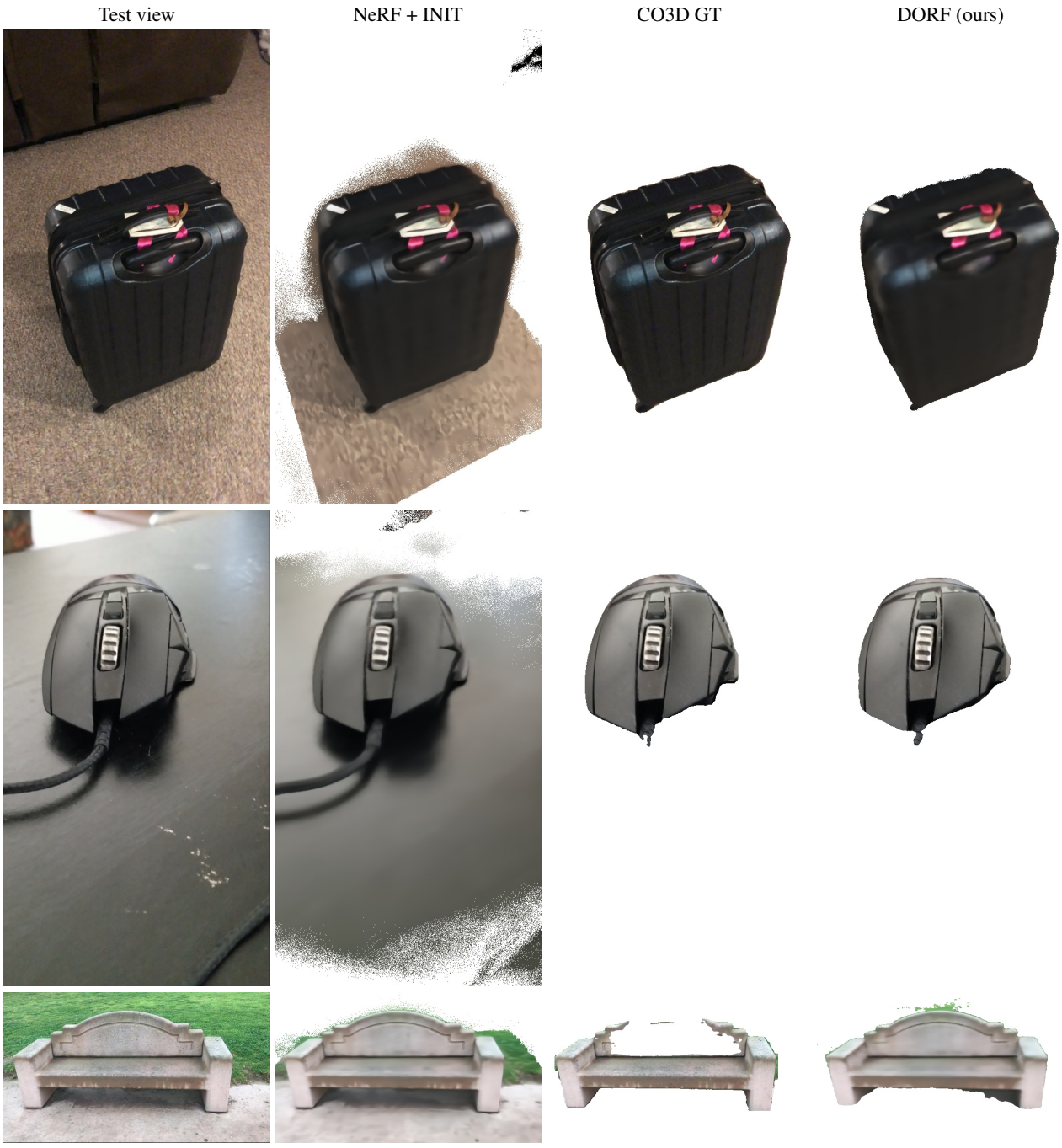


Figure 5. **Qualitative results of our reconstruction on different objects in CO3D [5]**. Our method can handle objects with various shape and appearance, and produce high quality reconstruction of the objects even with thin structures. The object masks by NeRF + INIT are often noisy and contain most of the background portion. Compared to the pesudo ground truth provided in CO3D [5] (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.

References

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

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

[3] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik,

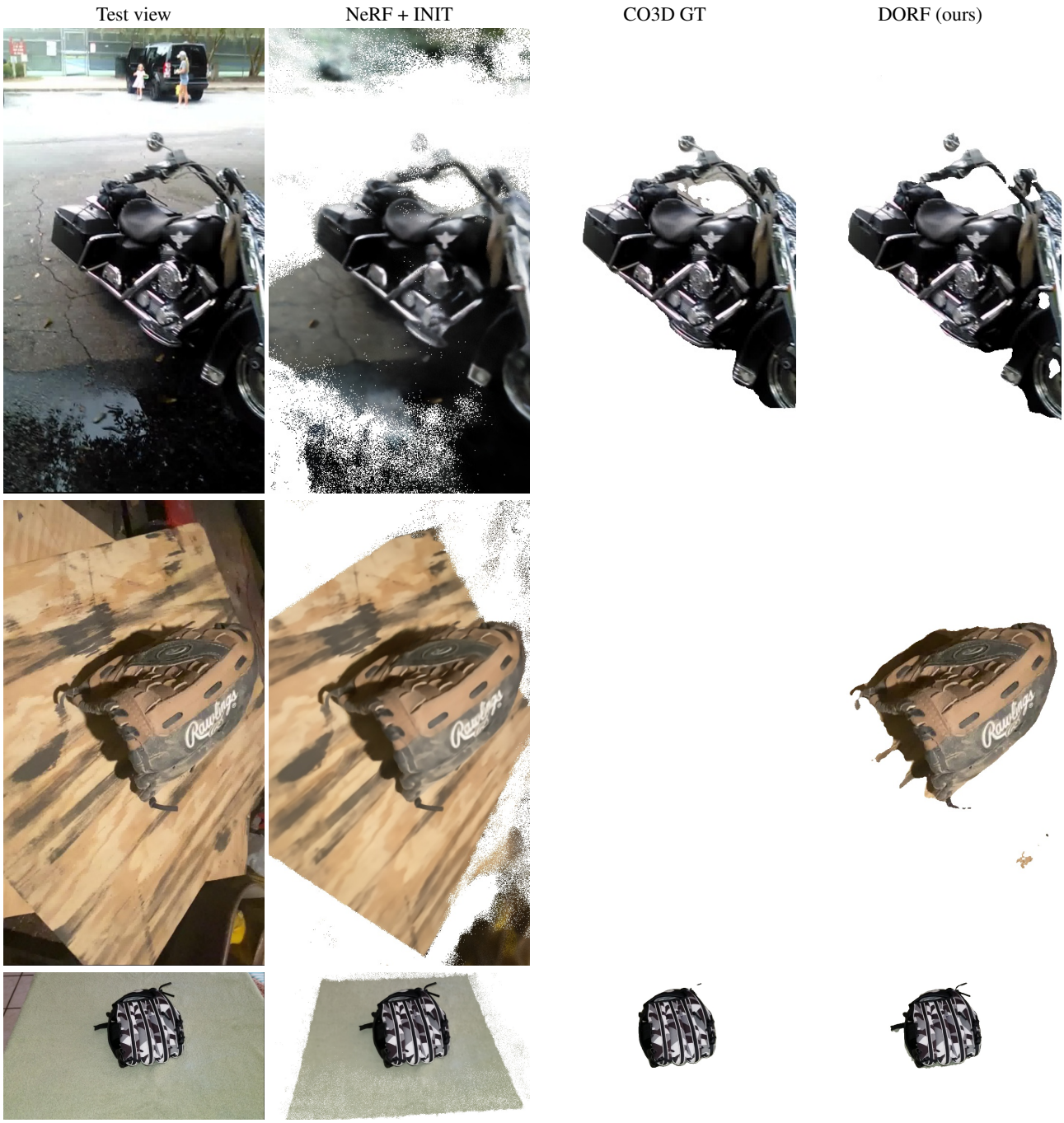


Figure 6. **Qualitative results of our reconstruction on different objects in CO3D [5]**. Our method can handle objects with various shape and appearance, and produce high quality reconstruction of the objects even with thin structures. The object masks by NeRF + INIT are often noisy and contain most of the background portion. Compared to the pesudo ground truth provided in CO3D [5] (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.

Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2, 3

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

[5] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Com-

Figure 7. **Qualitative results of our reconstruction on different objects in CO3D [5]**. 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 [5] (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.

mon objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *ICCV*, 2021. 1, 3, 4, 5, 6, 7

[6] 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

Figure 8. **Qualitative results of our reconstruction on different objects in CO3D [5]**. 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 [5] (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.