Supplementary Material

In this supplemental material, we provide the procedures for single object rendering and editing in Sec. A, details of our experiments in Sec. B, more experimental results in section C, limitations of our work in Sec. D, and finally potential social impacts in Sec. E. We gently urge readers to review our supplemental videos for dynamic visualization of qualitative results.

A Procedures for Single Object Rendering and Editing

After obtaining object-level segmentation, we can extract individual objects from the scene. Unlike conventional object saliency [5], we take 3D information into account. Single object rendering for object k is achieved by setting density to zero at coordinates unoccupied by object k, followed by performing typical volume rendering:

$$\sigma_k(\mathbf{x}) = m_k(\mathbf{x})\sigma(\mathbf{x}) \,, \tag{1}$$

$$\mathbf{C}_k(\mathbf{r}) = \sum_{p=1}^{P} \hat{T}(t_p)\, \alpha\left(\sigma_k(t_p)\delta_p\right) c(t_p) \,, \tag{2}$$

where $\hat{T}(t_p) = \exp\left(-\sum_{p'=1}^{p-1} \sigma_k(t_p)\delta_p\right)$, $\alpha\left(x\right) = 1 - \exp(-x)$, and $\delta_p = t_{p+1} - t_p$ is the distance between two adjacent quadrature sample points. Our system supports several 3D scene editing effects, similar to [14, 12]. We can implement arbitrary rigid transformation with each object. We can also achieve object insertion, duplication and removal as well. We show editing results in Sec. C.4 and the supplemental video.

B More Experimental Details

B.1 Inplementation of our system

We adopt TensoRF [1], an explicit grid version of NeRF [9] to build our system. We use the official implementation of TensoRF. We fix the number of voxels 2097156 when training, and use default hyperparameters. To estimate initial coarse label maps for computing the initial semantic estimation, we use IEM [11] for single object scenes, and DFC [7] with K-means clustering for multi-object scenes. We provide inplementation details of IEM and DFC in Sec. B.4. We implement our system in PyTorch and run all of our experiments with a single NVidia Tesla P40 GPU. Depending on the number of views and resolution of the scene, the training time ranges from 1 to 3 hours. Such reasonable time cost is mainly due to the architectural design of TensoRF. With our unsupervised method not involving any training of neural networks, and TensoRF based on voxel rather than MLP, the implementation of our system is **deep learning-free**. Our approach should also be able to be built upon other radiance field-based representations such as the vanilla NeRF [8] or plenoxel [4].

B.2 Datasets

We first test RFP on scenes with a single foreground object using two real datasets, **Local Light Field Fusion (LLFF)** [8] and **Common Objects in 3D (CO3D)** [10]. LLFF dataset contains real-world scenes captured with 20 to 62 roughly forward-facing images. All images are 1008×756 pixels. LLFF does not provide ground-truth label maps, and we only show qualitative results on the LLFF dataset without quantitative results. CO3D dataset contains real-world scenes captured with 50 to 120 images with various image resolutions. Some ground truth labels of CO3D are incomplete or noisy, and we exclude them when calculating quantitative metrics. To test RFP on scenes with multiple objects, we build a synthetic dataset upon ClevrTex [6], which allows wrapping realistic textures to the objects and background as well. Rather than limiting ourselves to regular geometric models, we use objects with more natural shapes, such as animals and furniture items. We use Blender [3] to render images and ground-truth label maps from 50 to 100 different viewpoints for each scene. Each scene of our synthetic dataset contains 2 to 4 objects. For all datasets, we use original resolution images as input. We split the datasets into training views and testing views with a ratio of around 9 to 1.

B.3 Metrics

We adopt the widely-used pixel classification accuracy (Acc.) and mean intersection over union (mIoU) as our metric. Acc. measures the proportion of pixels that have been assigned to the correct region. mIoU is the ratio between the area of the intersection between the inferred mask and the ground truth over the area of their union. Higher Acc. and mIoU mean better results. To evaluate scene segmentation in 3D, we report the metrics computed for both training and novel views. The former enables direct comparison to 2D methods, while the latter reflects the 3D nature. We denote Acc. and mIoU of novel views as N-Acc. and N-mIoU. Note that for our method, the difference between the training and the novel view is only related to whether the color image of the view is provided. The semantic label of any view is not input. Differently, for the supervised methods [15, 12], both the color image of the training view and the ground truth semantic label are provided as inputs.

B.4 Baselines

SemanticNeRF [15] and **ObjectNeRF** [12] are supervised approaches as they take annotated labels as input to supervise a semantic branch for object separation. We feed the ground truth labels as input to these methods where the official implementations are used. **uORF** [13] is an approach for unsupervised object discovery pretrained on large class-specific datasets. We use the official implementation and the weights pretrained using the Room-Diverse dataset. **IEM** [11] is a single image-based unsupervised segmentation approach taking one unlabeled image as input and does not involve any external data or network training. We also use its official implementation. We resize input images to 300x300 and set all the other hyperparameters to default values. **ReDO** [2] is a single image-based unsupervised segmentation approach using generative models. For each scene, we train the model using all the images. Again the official implementation and default hyperparameters are used. **DFC** [7] is an unsupervised image segmentation approach that can partition an image into several areas corresponding to different objects, and thus is different from IEM and ReDO which can only perform foreground-background partitioning. DFC requires scribbles or given foreground-background separation. Thus, we first feed each input image into a deep feature extractor and cluster all the pixels into two classes using k-means clustering. Our implementation of DFC is used.

C More Experimental Results

C.1 3D Segmentation of Scenes with a Single Foreground Object

Fig. 1 shows more qualitative comparison with IEM [11] on the LLFF [8] dataset; Figs 2 and 3 show qualitative comparison on the CO3D [10] dataset. Please also refer to the supplemental video.

C.2 3D Segmentation of Scenes with Multiple Objects

Fig. 4 shows more qualitative comparison with DFC [7] on the RFP synthetic dataset. Please also refer to the supplemental video.

C.3 Ablation Study

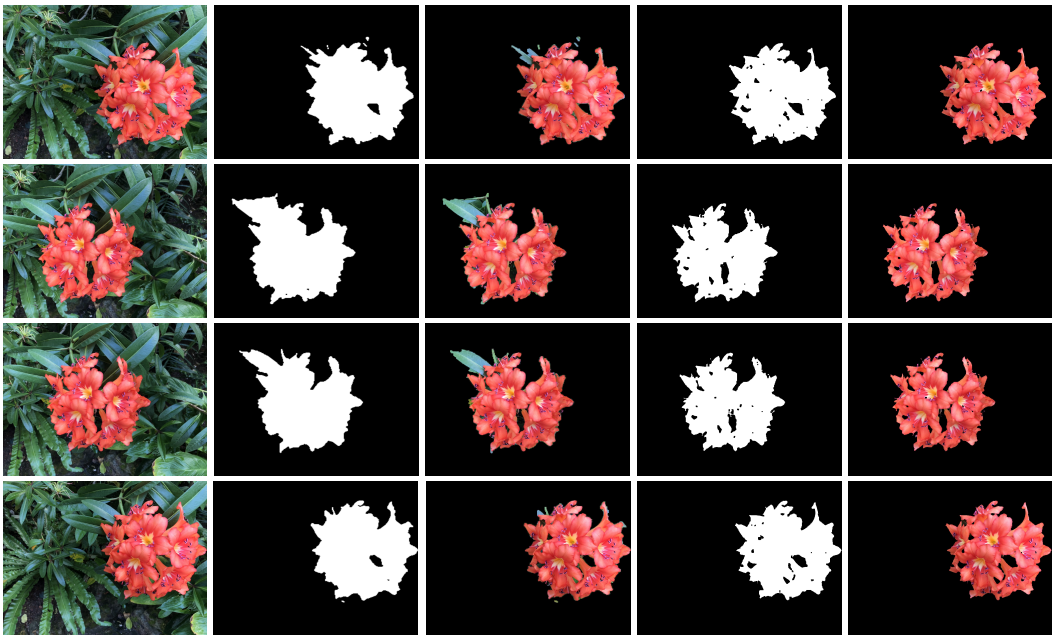
We show more qualitative evaluation of different components of our approach in Fig. 5 and Fig. 6. Please also review the supplemental video.

C.4 Scene Editing

Fig. 7 shows more qualitative results of scene editing. Please also review the supplemental video. Our approach cannot correctly render objects parts in total occlusion, which can be remedied by the 3D guard mask described in [12]. Moreover, our editing results can estimate wrong shading for the translated/rotated objects since the object radiance fields are learnt under the given illumination.

C.5 Novel view synthesis

In this subsection we show quantitative comparison on the task of novel view synthesis between our method and other radiance field-based scene representations. Although not the focus of this work,



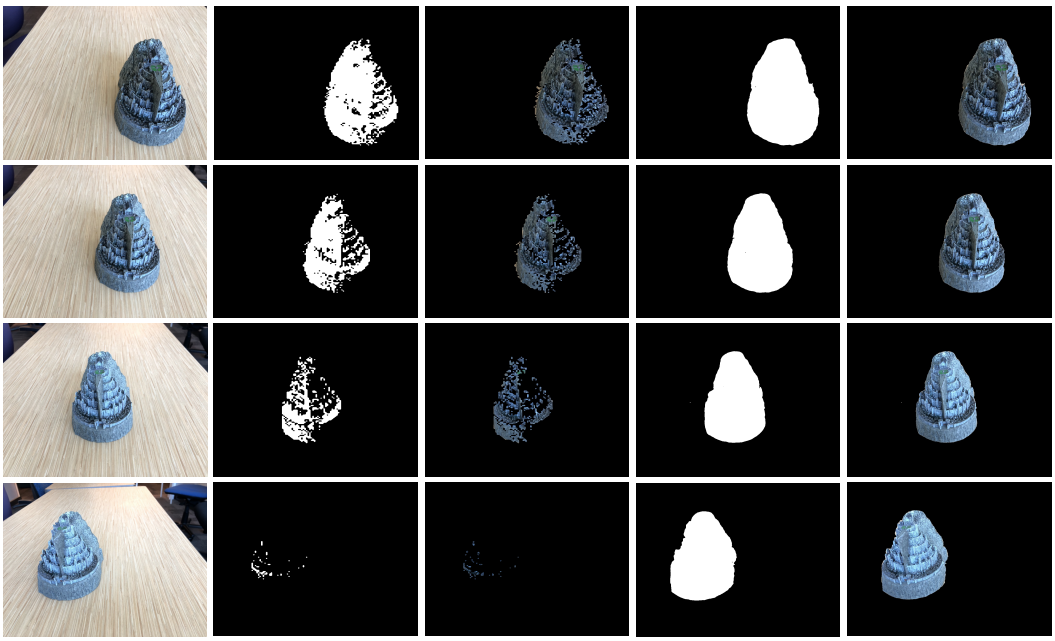


Figure 1: Qualitative comparison on the LLFF [8] dataset.

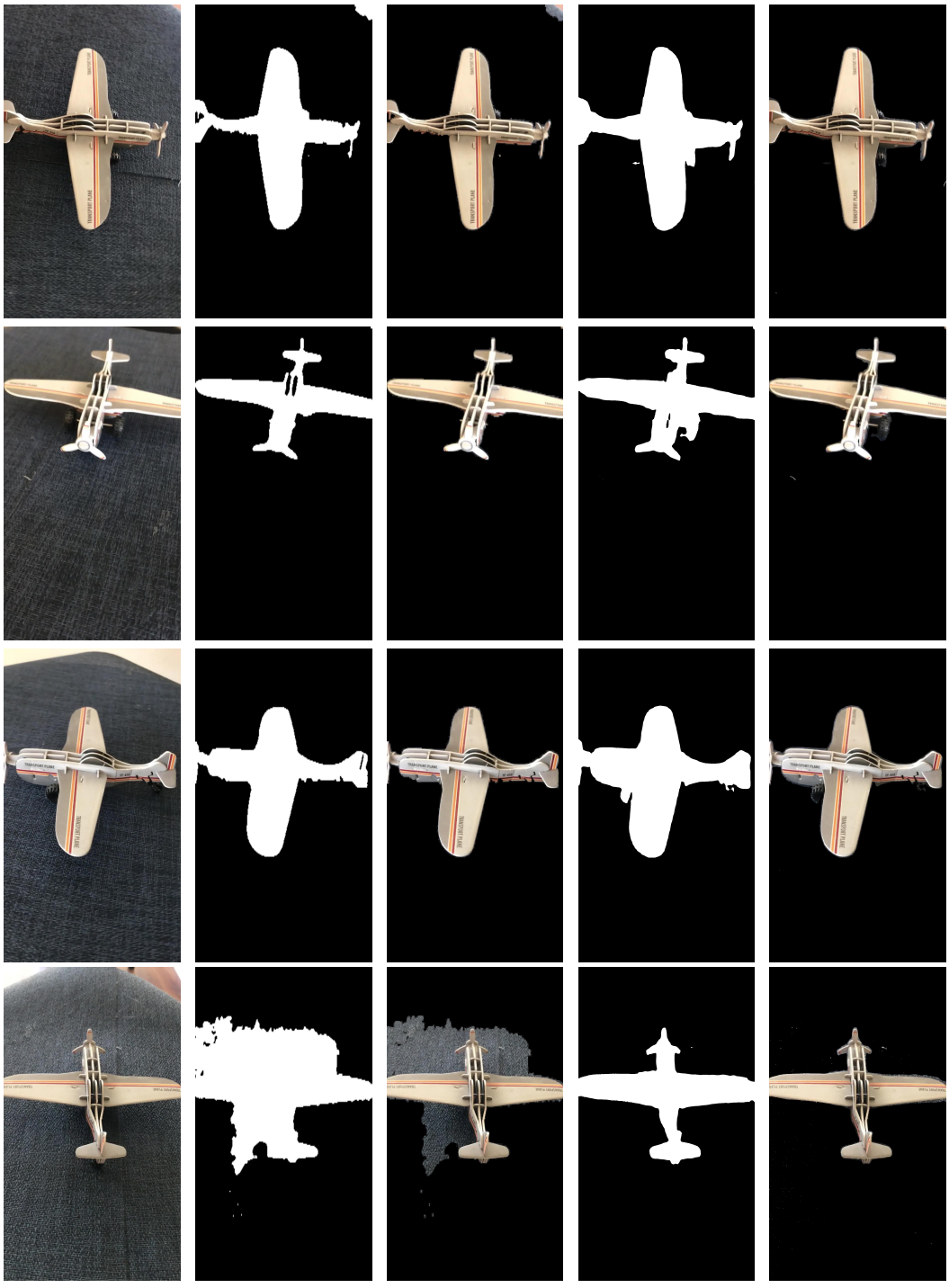


Figure 2: Qualitative comparison on the CO3D [10] dataset.

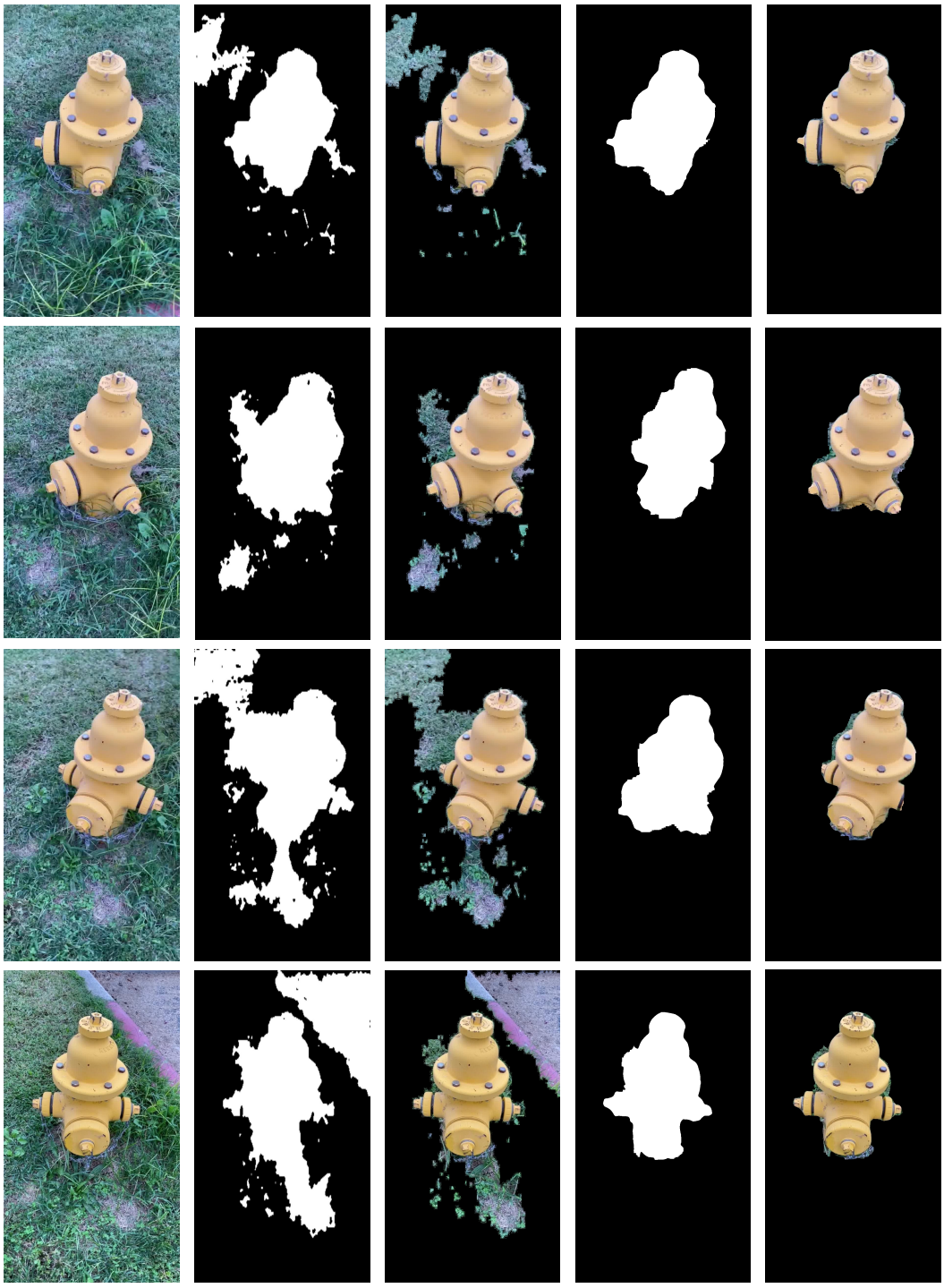


Figure 3: Qualitative comparison on the CO3D [10] dataset.

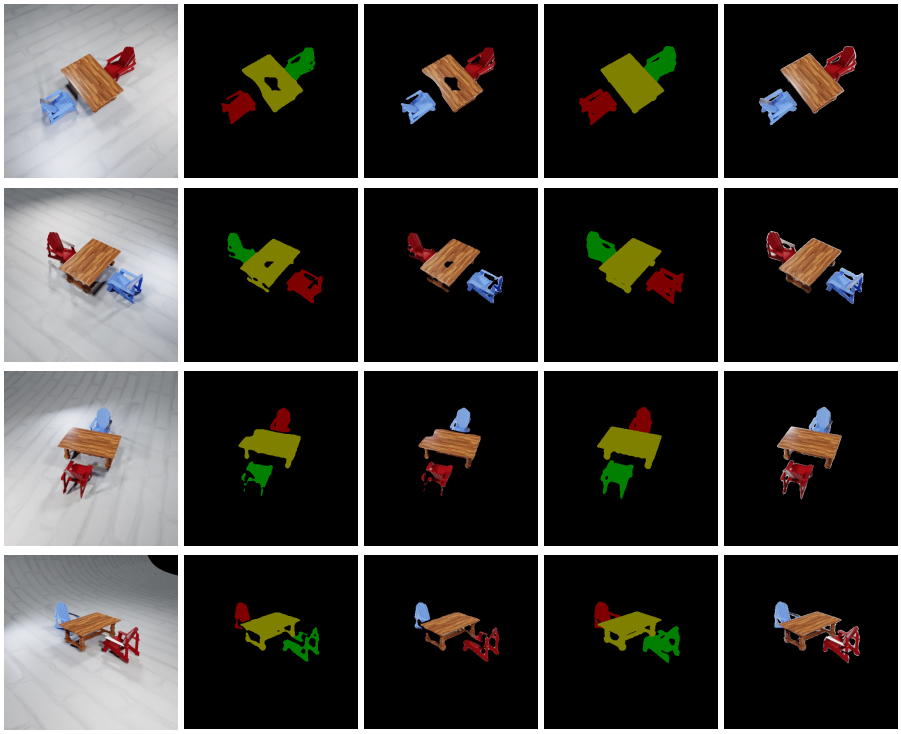


Figure 4: Qualitative comparison on the RFP synthetic dataset.

Figure 5: Qualitative ablation study on the CO3D [10] dataset.

Methods	PSNR↑	SSIM↑
NeRF [9]	26.50	0.811
TensoRF [1]	26.73	0.839
RFP (ours)	26.41	0.807

Table 1: Quantitative comparison on novel view synthesis.

this experiment is to show that our method does not sacrifice realism of rendering. We compare our approach with the vanilla NeRF [9] and TensoRF [1]. Tab. 1 tabulates PSNRs and SSIMs on the LLFF [8] dataset. The reported values shows a small gap between our method and other methods, where the gap in rendering realism should be reduced by more custom hyperparameters and longer training.

D Limitations

As the first significant attempt in enabling unsupervised multi-view object segmentation on NeRF, our approach has limitations which will lead to fruitful future research. First, we perform propagation upon the color fields of individual objects while the density field can also be better exploited. While this approach may sound naive, how to take advantage of the relationship between the density field, the semantic field, and the color fields, or even "inpaint" the density field is worth exploring. Second,

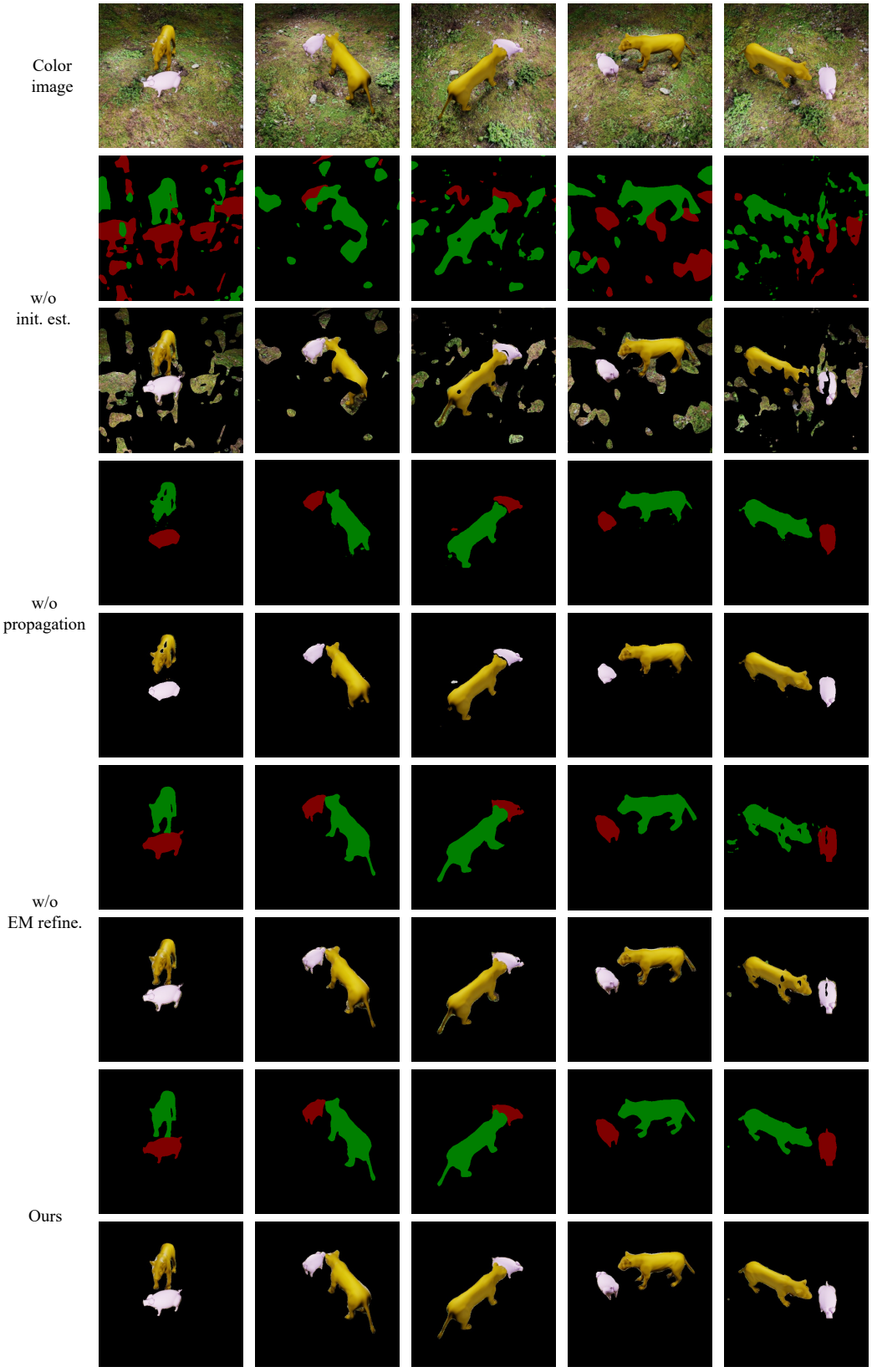


Figure 6: Qualitative ablation study on the RFP synthetic dataset.

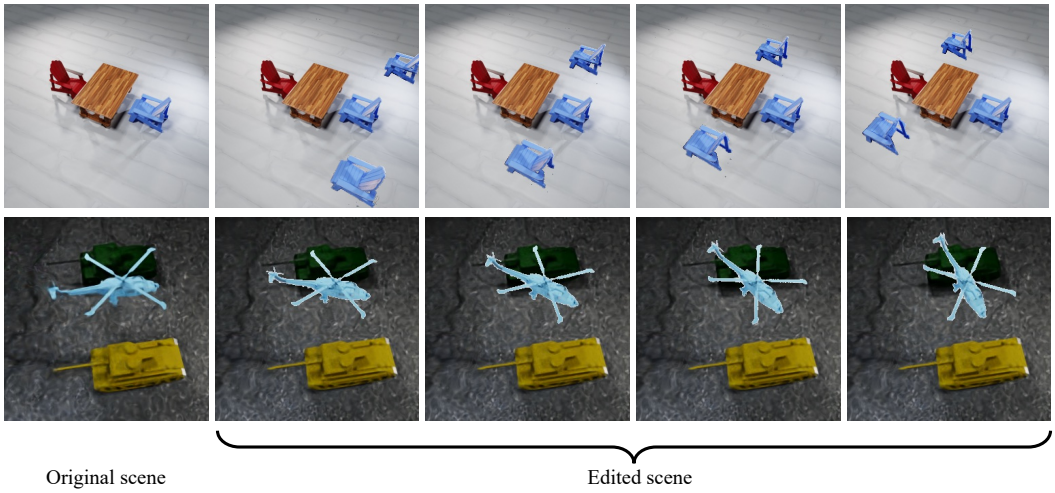


Figure 7: Scene editing effects.

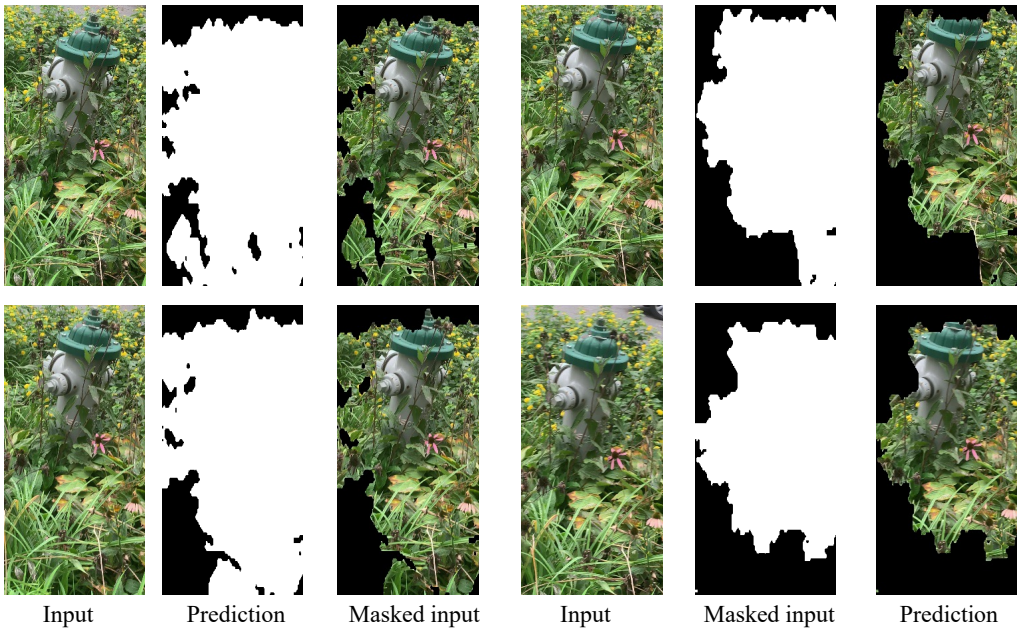


Figure 8: Failure cases of our methods.

our method cannot handle cases when the background is too complex, or when the background and the foreground have similar colors which are hard to distinguish. We show some failure cases in Fig. 8. Finally, while achieving individual object rendering and editing, our approach cannot correctly render objects parts in total occlusion. Yet this drawback can be remedied by the 3D guard mask described in [12]. Moreover, our editing results can estimate wrong shading for the translated/rotated objects since the object radiance fields are learned under the given illumination.

E Potential Negative Societal Impacts

Misuse of scene editing, such as generating images with real people, poses a societal threat, and we do not condone using our work with the intent of spreading misinformation or tarnishing reputation.

References

[1] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision*, 2022.

[2] Mickaël Chen, Thierry Artières, and Ludovic Denoyer. Unsupervised object segmentation by redrawing. In *Advances in Neural Information Processing Systems*, 2019.

[3] Blender Online Community. *Blender – a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018.

[4] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *International Conference on Computer Vision*, pages 5501–5510, 2022.

[5] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, and Shipeng Li. Salient object detection: A discriminative regional feature integration approach. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference*, 2013.

[6] Laurynas Karazija, Iro Laina, and Christian Rupprecht. Clevrtex: A texture-rich benchmark for unsupervised multi-object segmentation. *arXiv preprint arXiv:2111.10265*, 2021.

[7] Wonjik Kim, Asako Kanezaki, and Masayuki Tanaka. Unsupervised learning of image segmentation based on differentiable feature clustering. *IEEE Transactions on Image Processing*, 29:8055–8068, 2020.

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

[9] 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 *European Conference on Computer Vision*, 2020.

[10] 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 *International Conference on Computer Vision*, pages 10901–10911, 2021.

[11] Pedro Savarese, Sunnie SY Kim, Michael Maire, Greg Shakhnarovich, and David McAllester. Information-theoretic segmentation by inpainting error maximization. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference*, pages 4029–4039, 2021.

[12] 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 *International Conference on Computer Vision*, 2021.

[13] Hong-Xing Yu, Leonidas J. Guibas, and Jiajun Wu. Unsupervised Discovery of Object Radiance Fields. In *International Conference on Learning Representations*, 2021.

[14] Jiakai Zhang, Xinhang Liu, Xinyi Ye, Fuqiang Zhao, Yanshun Zhang, Minye Wu, Yingliang Zhang, Lan Xu, and Jingyi Yu. Editable free-viewpoint video using a layered neural representation. *ACM Transactions on Graphics (TOG)*, 40(4):1–18, 2021.

[15] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *International Conference on Computer Vision*, pages 15838–15847, 2021.