Volumetric Disentanglement for 3D Scene Manipulation

Sagie Benaim[1] Frederik Warburg[2,*] Peter Ebert Christensen[1,*] Serge Belongie[1]
[1]University of Copenhagen [2]Technical University of Denmark

1. Training and Implementation Details

For training, we consider the natural non-synthetic scenes given in [1], together with their associated pose information. An off-the-shelf segmentation or manual annotation is used to extract masks. We note that masks need not be exact, and may capture more then the desired object (see main paper for details). Our rendering resolution for training the background and full scenes is 504×378. For the manipulation tasks, the same resolution is used for *3D inpainting*, *object camouflage*, *transformation* and *non-negative inpainting* tasks. For the *semantic manipulation* task, our rendering resolution is 252×189. For the CLIP [2] input, for a given view, we sample a 128×128 grid of points from the 252×189 output, and then upsample it to 224×224, which is the required input resolution of CLIP. We normalize the images and apply a text and image embedding as in CLIP [2]. We follow NeRF [1], in optimizing both a "coarse" and "fine" networks for a neural radiance field, and follow the same sampling strategy of points along the ray. All neural fields are parametrized using an MLP with ReLU activation of the same architecture of [1]. We use an Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with a learning rate that begins with 5×10^{-4} and decays exponentially to 5×10^{-5}.

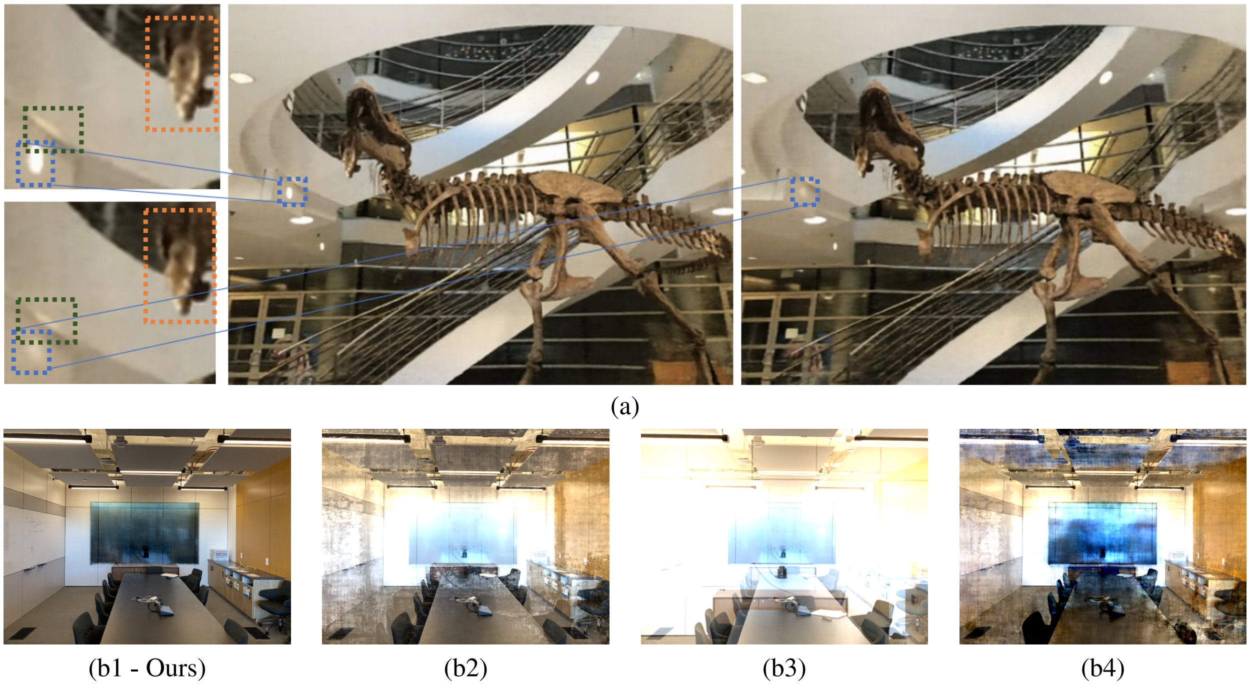
2. Additional Visualizations

As noted in the main text, Fig. 1 (a) shows the failure to remove a light source. In Fig. 1 (b1 to b4), we show, for the task of foreground object translation (Fig. 6), alternatives to the recombining method of Eq. 5, with (b2) ${c'_{full}}^{i_r}$ instead of ${c'_{fg}}^{i_r}$, (b3) ${w'_{full}}^{i_r}$ instead of ${w'_{fg}}^{i_r}$, (b4) $c^c_r = \sum_{i=1}^{N} ({w'_{bg}}^{i_r} + {w'_{fg}}^{i_r}) \cdot ({c'_{bg}}^{i_r} + {c'_{fg}}^{i_r})$.

3. Training masks

We provide a sample of the training masks used for training views in Fig. 2.

[*] Contributed equally.



(a)

(b1 - Ours)

(b2)

(b3)

(b4)

Figure 1. (a) *Failure to completely remove a light source.* The original light source is shown in blue in the middle image and for the background, using our method, on the right. In orange and green are regions affected by the light source, resulting in the failure to completely remove it. Full 3D scene is shown in the supplementary HTML. (b1-b4) *Ablation for composition.* Alternatives to the composition shown in Eq. 5 for foreground object translation (Fig. 6).

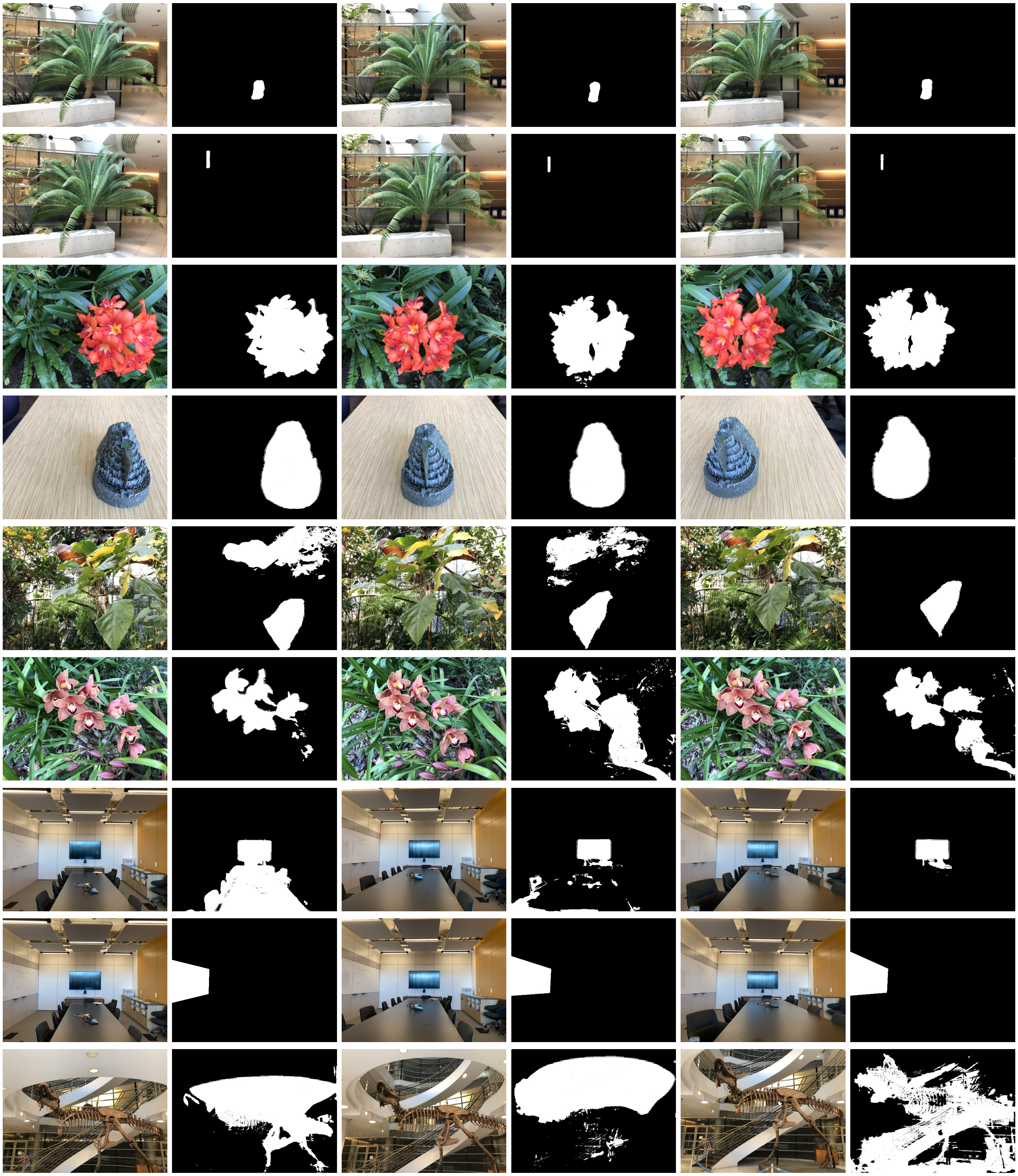


Figure 2. Sample of the masks used for our method for training views.

References

[1] Mildenhall, B., Srinivasan, P., Tancik, M., Barron, J., Ng, R.: Representing scenes as neural radiance fields for view synthesis. In: Proc. of European Conference on Computer Vision, Virtual (2020) 1

[2] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021) 1