# Supplemental Material of CLIP-NeRF: Text-and-Image Driven Manipulation of Neural Radiance Fields

Can Wang[1], Menglei Chai[2], Mingming He[3], Dongdong Chen[4], Jing Liao[1,†]
[1]City University of Hong Kong [2]Snap Inc. [3]USC Institute for Creative Technologies [4]Microsoft Cloud AI
cwang355-c@cityu.edu.hk, cmlatsim@gmail.com, hmm.lillian@gmail.com
cddlyf@gmail.com, jingliao@cityu.edu.hk


Figure 1. Continuous manipulation results. Our method supports editing both the shape and appearance with a single text prompt or an exemplar by continuously editing the shape and appearance.

## 1. Extended Discussions

**Continuous Manipulation.** Our method supports editing both shape and appearance, given a single text prompt or an examplar. This can be achieved by continuously editing the shape and appearance, i.e., first editing the shape and then editing the color and vice versa. We show results in Fig. 1. This provides a user-friendly way for editing when users want to edit both the shape and appearance indicated by a single text description or an exemplar.

**Fine-grained appearance manipulation within a same color category**. Though our method cannot handle fine-grained local parts shape and appearance edits as stated in the limitation, it supports fine-grained appearance manipulation at a whole object level, as shown in Fig. 2. Our method enjoys achieving various editing results within a same color category. Without loss of generality, we show various editing results related to the color blue.


Figure 2. Fine-grained appearance manipulation results within a same color category.

**Scaling along Editing Direction.** From equation 1, our code mappers provide manipulation directions $\Delta z_s = \big(\hat{\mathcal{E}}_t(\boldsymbol{t})\big)$ and $\Delta z_a = \mathcal{M}_a\big(\hat{\mathcal{E}}_t(\boldsymbol{t})\big)$ in the latent space for shape and appearance editing. We can scale along the editing direction to obtain gradually editing results through the following equation:

$$\begin{aligned} \boldsymbol{z}_s &= s \times \Delta z_s + \boldsymbol{z}'_s, \\ \boldsymbol{z}_a &= s \times \Delta z_a + \boldsymbol{z}'_a, \end{aligned} \tag{1}$$

where $s$ is the scalar. This scaled scheme also supports directions learned from examplars. We show scaled manipulation results in Fig. 3. The manipulation effect becomes stronger as the scalar $s$ increases.

**Interpolation.** As shown in Fig. 4, the shape and appearance latent space supports interpolation between two latent codes $\boldsymbol{z}^1 = (\boldsymbol{z}_s^1, \boldsymbol{z}_a^1)$ and $\boldsymbol{z}^2 = (\boldsymbol{z}_s^2, \boldsymbol{z}_a^2)$. Given an interpolation ratio $\boldsymbol{r}$, we define the interpolated latent code $\boldsymbol{z}_{inter}$

*† Jing Liao is the corresponding author. Our project page is https://cassiepython.github.io/clipnerf/

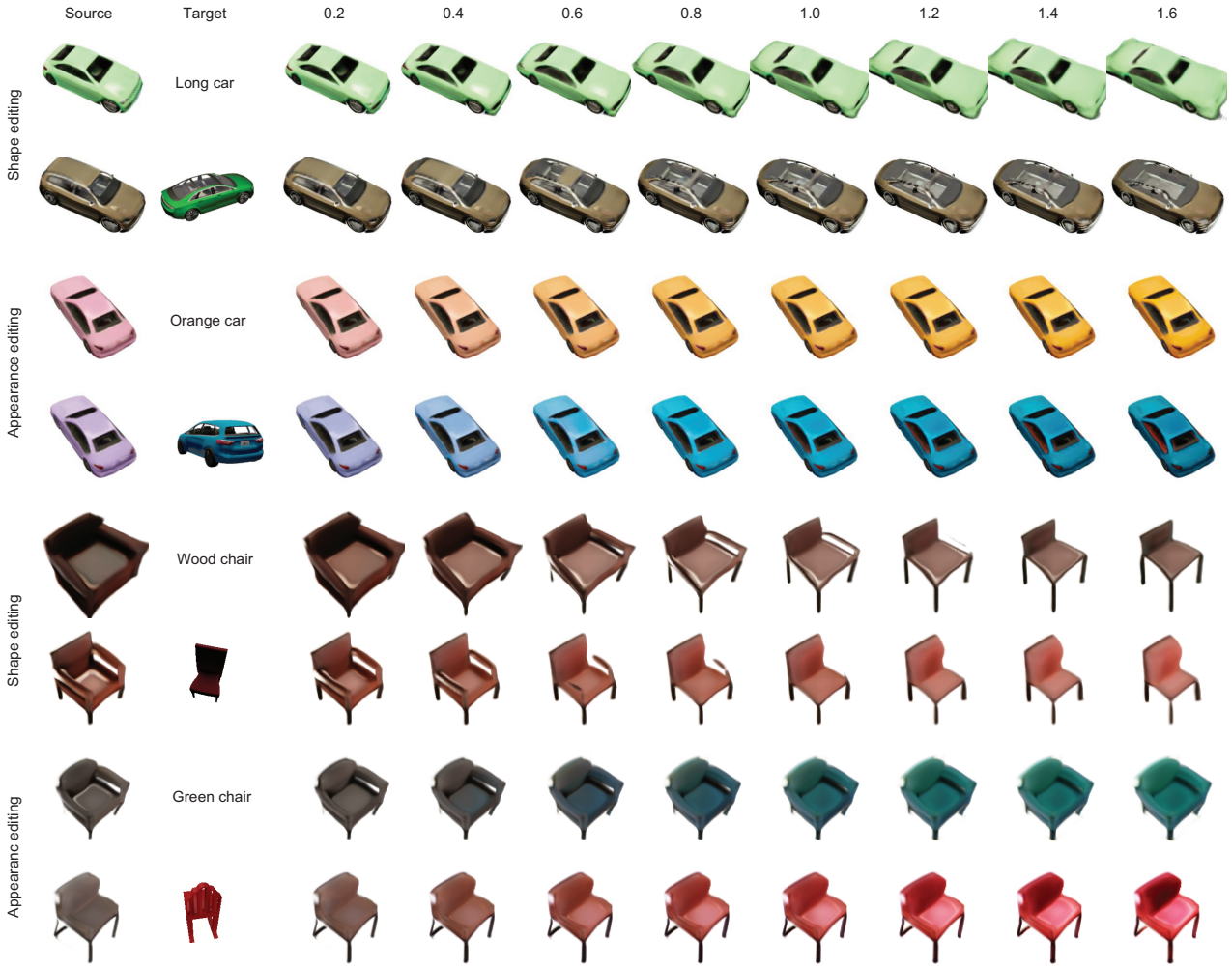


Figure 3. Editing results of moving along the $\Delta \boldsymbol{z}_s$ and $\Delta \boldsymbol{z}_a$ directions. The shape and appearance codes of the source are updated by adding $s * \Delta \boldsymbol{z}_s$ and $s * \Delta \boldsymbol{z}_a$, while $s$ is a scalar ranging from 0 to 1.6 with a step 0.2.

as $\boldsymbol{z}_{inter} = \boldsymbol{z}^2 \times \boldsymbol{r} + \boldsymbol{z}^1 \times (1 - \boldsymbol{r})$, while $\boldsymbol{r}$ ranges from 0 to 1.0 with a step 0.1. Then we can obtain the interpolated result using $\boldsymbol{z}_{inter}$.

**Necessity of latent space.** Our method performs shape and appearance edits on the latent space of a conditional NeRF model with our designed CLIP constraints. A question arises whether the latent space is necessary, i.e., is it possible to edit the shape and appearance of a single NeRF model [2] directly rather than a conditional NeRF? We first evaluate our designed CLIP loss on appearance editing in Fig. 5. Given a pre-trained NeRF model on the LLFF dataset [1], we fix the density-related layers and finetune the color-related layers of NeRF with our CLIP loss. We also use the patch-based ray samplar while calculating our CLIP loss. Our CLIP constraint succeeds in editing the color of a single NeRF model without any ground truth. However, we fail to achieve satisfying results while editing the shape of a single NeRF with our deformation network conditioned by a text prompt. This may be because the CLIP loss is still not strong and compact enough to deform the shape without the latent space constraint. We think it is an interesting problem to explore in the future.

## 2. Supplementary Video

We provide a supplementary video with a real-time demo and more visual results rendered in multiple views. We highly recommend watching our supplementary video to observe the user-friendliness and view-consistency that our method can achieve in both shape and color editing.

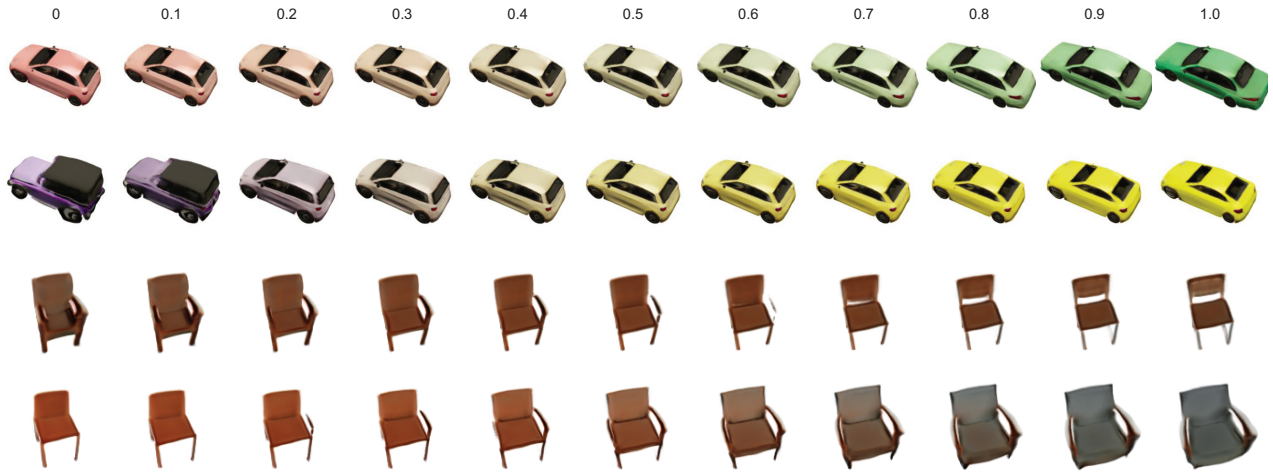


Figure 4. Interpolations in the latent space. The leftmost and rightmost ones are the inputs. And interpolation ratios are shown at the top.

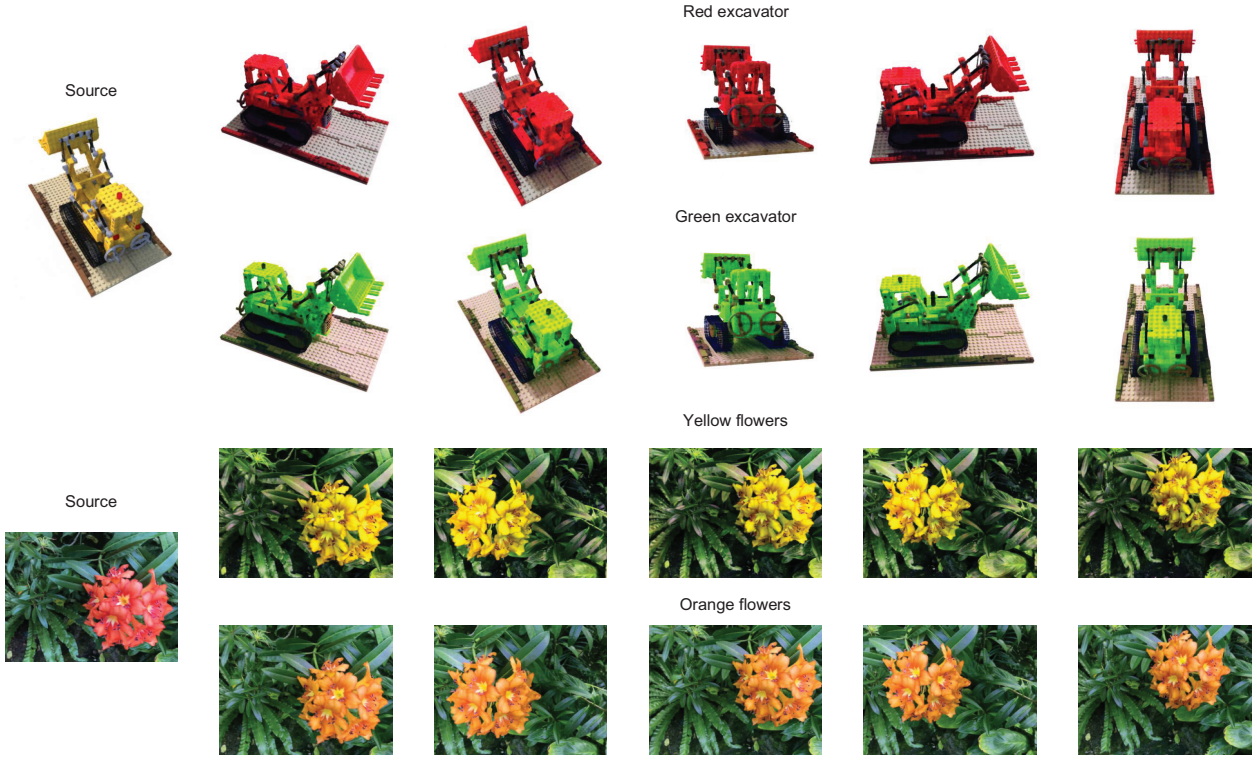


Figure 5. Single NeRF appearance editing results with our designed CLIP loss. NeRF models are trained on LLFF dataset [1].

## References

[1] 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. 2, 3

[2] 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*, pages 405–421. Springer, 2020. 2