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

2022.03.14
Presenter: Junha Hyung

Introduction

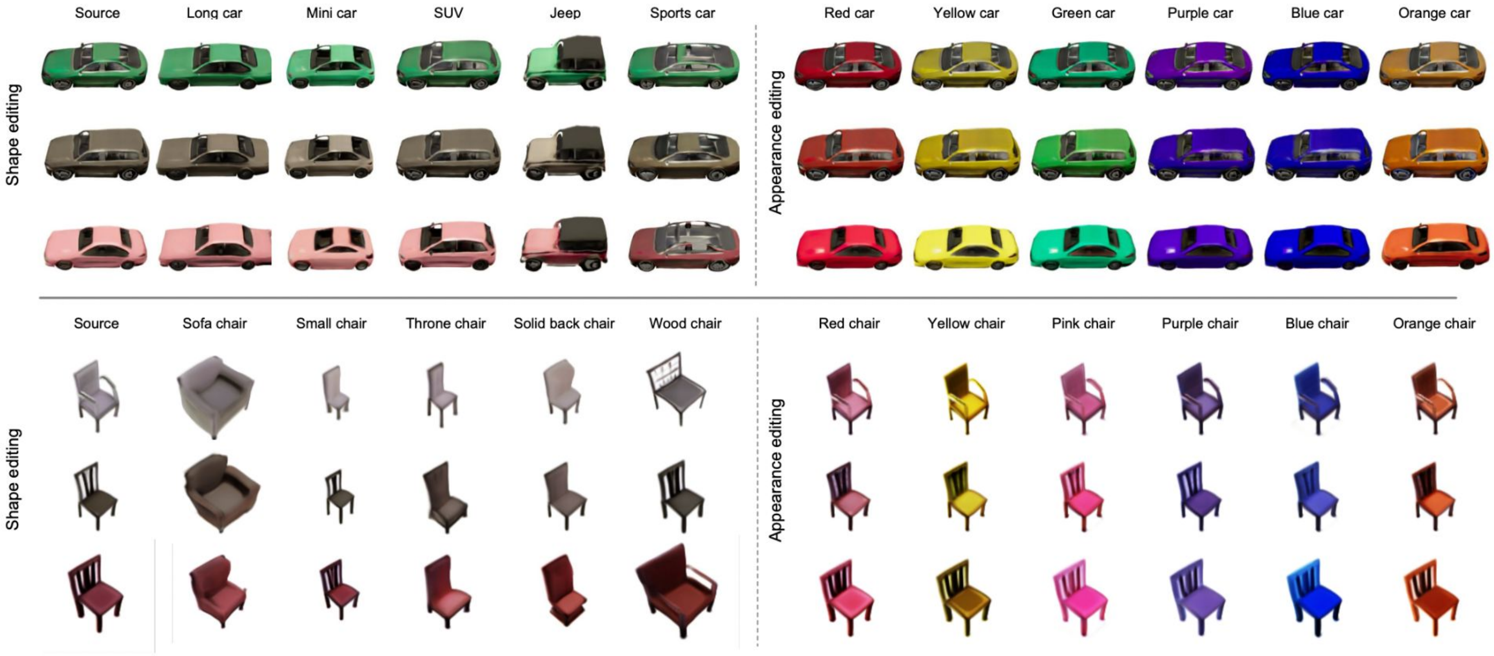


Figure 6. **Text-Driven Editing Results.**

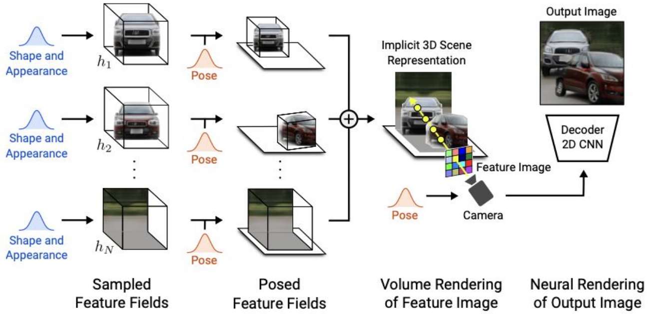
Contribution

- The first text-and-image-driven manipulation method for NeRF, using a unified framework to provide users with flexible control over 3D content using either a text prompt or an exemplar image.
- A disentangled conditional NeRF architecture by introducing a shape code to deform the volumetric field and an appearance code to control the emitted colors.
- Feedforward code mappers that enable the fast inference

Related works

Generative Neural Radiance Field

- object와 background의 shape/appearance code를 sampling해서, arbitrary한 pose에 따라 image들을 rendering.
- 기존 2D-GANs와는 달리, camera pose를 조작할 수 있음.



▲ GIRAFFE[1]

Edit NeRF

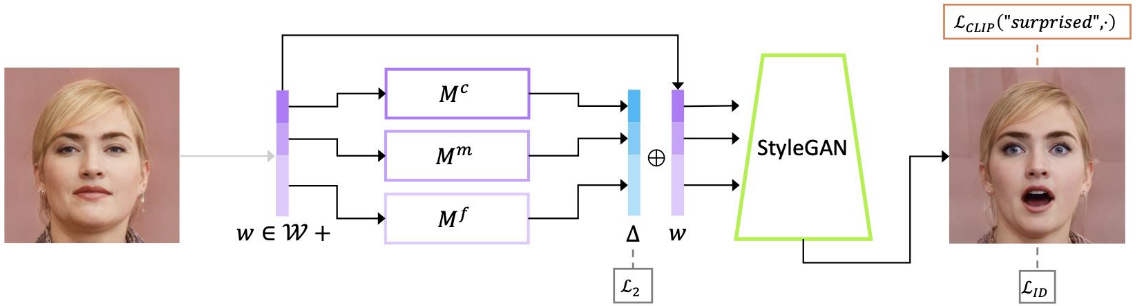
- Scribble로 color, shape editing
- Optimization based

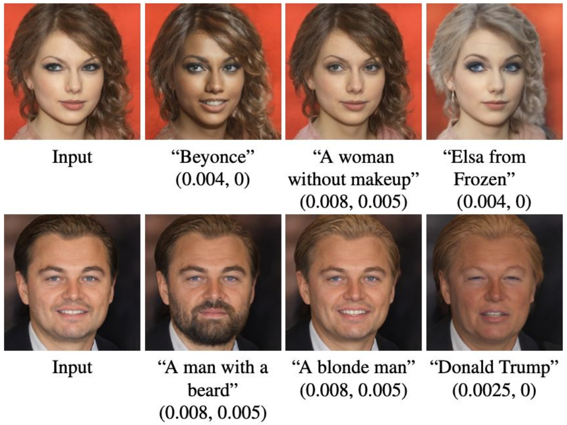
[1] GIRAFFE: Representing Scenes as Compositional Generative Neural Feature Fields. *CVPR*, 2021

Related works

Style Clip [1]

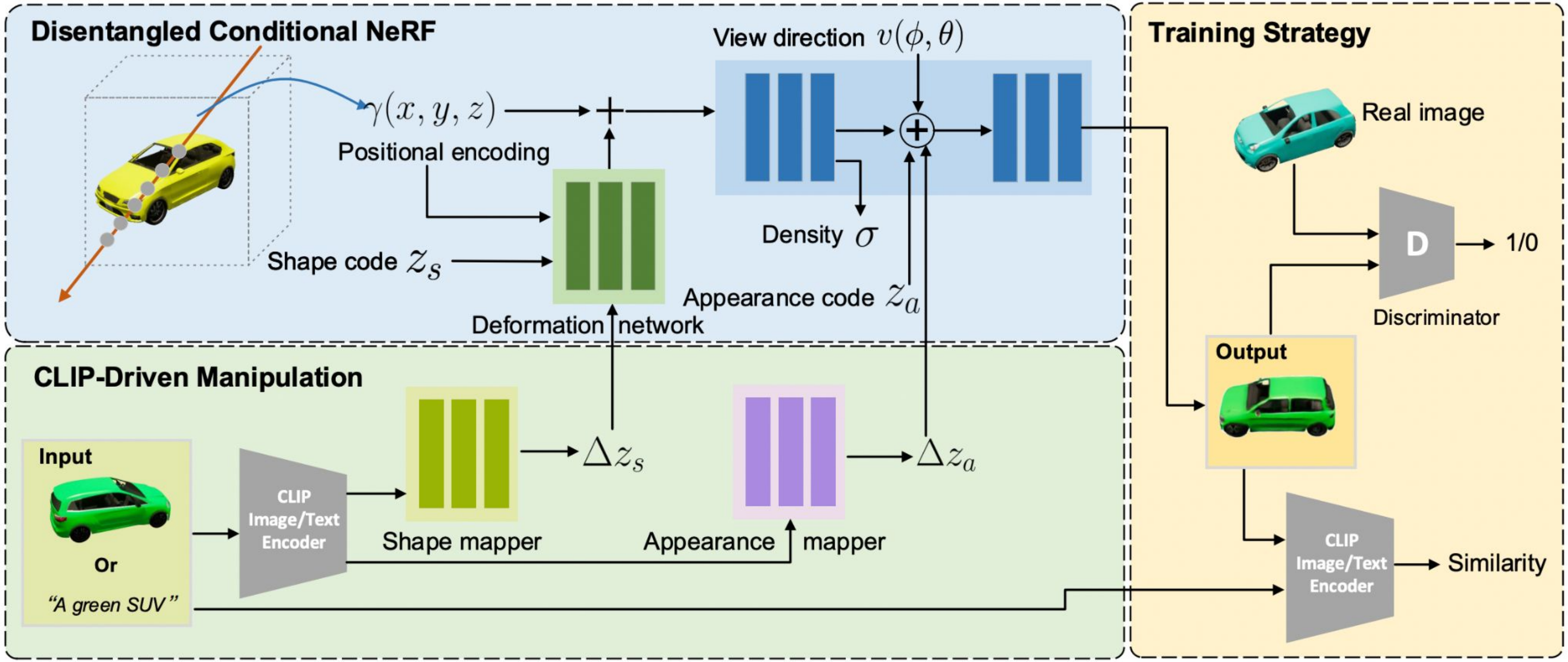
- Pretrained StyleGAN과, Clip 활용해 text guided image editing





[1] StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery

Method



Disentangled Conditional NeRF
View direction $v(\phi, \theta)$
$\gamma(x, y, z)$
Positional encoding
Shape code z_s
Density σ
Appearance code z_a
Deformation network
Training Strategy
Real image
D
1/0
Discriminator
Output
CLIP-Driven Manipulation
Input
Or
"A green SUV"
CLIP Image/Text Encoder
Shape mapper
Δz_s
Appearance mapper
Δz_a
CLIP Image/Text Encoder
Similarity

Overall formulation

$$\mathcal{F}'_\theta(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{z}_s, \boldsymbol{z}_a) : \big(\mathit{\Gamma}(\boldsymbol{x}) \oplus \boldsymbol{z}_s, \mathit{\Gamma}(\boldsymbol{v}) \oplus \boldsymbol{z}_a\big) \rightarrow (\mathbf{c}, \sigma), \quad (1)$$

Positional encoding

$$\gamma(p)_k = \begin{cases} \sin(2^k \pi p), & \text{if } k \text{ is even,} \\ \cos(2^k \pi p), & \text{if } k \text{ is odd,} \end{cases}$$

Conditional shape deformation

- Completely isolates the shape condition from affecting the appearance
- Regularize the output shape to be smooth deformation of the base shape

$$\gamma^*(p, \Delta p)_k = \gamma(p)_k + \tanh(\Delta p_k),$$

$$p \in \boldsymbol{p}, \Delta p \in \mathcal{T}(\boldsymbol{p}, \boldsymbol{z}_s)$$

Clip mapper

$$\boldsymbol{z}_s = \mathcal{M}_s\big(\hat{\mathcal{E}}_t(\boldsymbol{t})\big) + \boldsymbol{z}'_s,$$

$$\boldsymbol{z}_a = \mathcal{M}_a\big(\hat{\mathcal{E}}_t(\boldsymbol{t})\big) + \boldsymbol{z}'_a,$$

Training Strategy

- Two phase training
  - First train generative NeRF
  - Second freeze NeRF and train mapper networks

$$\mathcal{L}_{\text{GAN}} = \mathbb{E}_{\boldsymbol{z}_s \sim \mathcal{Z}_s, \boldsymbol{z}_a \sim \mathcal{Z}_a, \boldsymbol{v} \sim \mathcal{Z}_v} \left[f\left(\mathcal{D}(\mathcal{F}_\theta(\boldsymbol{v}, \boldsymbol{z}_s, \boldsymbol{z}_a)) \right) \right] + \mathbb{E}_{\mathbf{I} \sim d} \left[f\left(-\mathcal{D}(\mathbf{I}) + \lambda_r \|\nabla \mathcal{D}(\mathbf{I})\|^2 \right) \right]. \quad (7)$$

$$D_{\text{CLIP}}(\mathbf{I}, \boldsymbol{t}) = 1 - \left\langle \hat{\mathcal{E}}_i(\mathbf{I}), \hat{\mathcal{E}}_t(\boldsymbol{t}) \right\rangle,$$

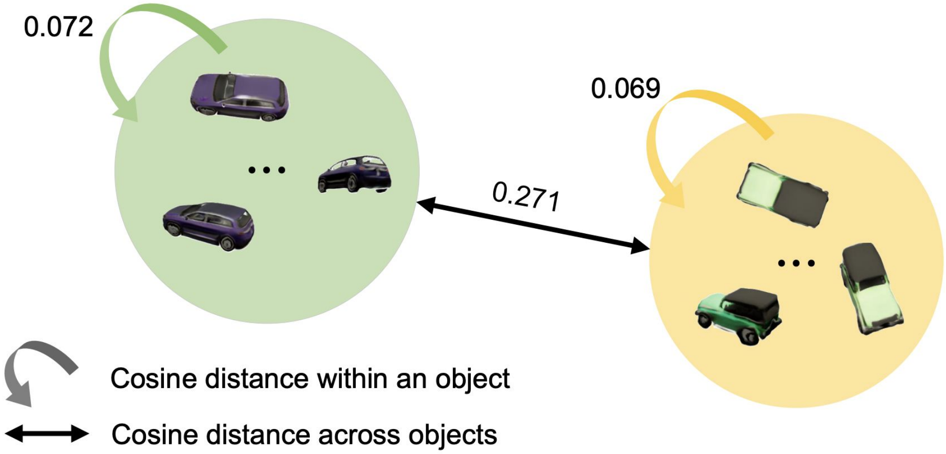
$$\mathcal{L}_{\text{shape}} = f\left(\hat{\mathcal{D}}\left(\hat{\mathcal{F}}_\theta\left(\boldsymbol{v}, \mathcal{M}_s(\hat{\mathcal{E}}_t(\boldsymbol{t})) + \boldsymbol{z}_s, \boldsymbol{z}_a \right) \right) \right) + \lambda_c D_{\text{CLIP}}\left(\hat{\mathcal{F}}_\theta\left(\boldsymbol{v}, \mathcal{M}_s(\hat{\mathcal{E}}_t(\boldsymbol{t})) + \boldsymbol{z}_s, \boldsymbol{z}_a \right), \boldsymbol{t} \right), \quad (8)$$

$$\mathcal{L}_{\text{appear}} = f\left(\hat{\mathcal{D}}\left(\hat{\mathcal{F}}_\theta\left(\boldsymbol{v}, \boldsymbol{z}_s, \mathcal{M}_a(\hat{\mathcal{E}}_t(\boldsymbol{t})) + \boldsymbol{z}_a \right) \right) \right) + \lambda_c D_{\text{CLIP}}\left(\hat{\mathcal{F}}_\theta\left(\boldsymbol{v}, \boldsymbol{z}_s, \mathcal{M}_a(\hat{\mathcal{E}}_t(\boldsymbol{t})) + \boldsymbol{z}_a \right), \boldsymbol{t} \right). \quad (9)$$

Discussion

Image editing을 하기에 Clip이 적절한 모델인가?

- Higher cosine similarity within an object



Inverse Manipulation

To be specific, during each iteration, we first optimize $\boldsymbol{v}$ while keeping $\boldsymbol{z}_s$ and $\boldsymbol{z}_a$ fixed using the following loss:

$$\begin{aligned}\mathcal{L}_v = &\left\|\hat{\mathcal{F}}_\theta(\boldsymbol{v}, \hat{\boldsymbol{z}_s}, \hat{\boldsymbol{z}_a}) - \mathbf{I}_r\right\|_2 + \\ &\lambda_v D_{\text{CLIP}}\left(\hat{\mathcal{F}}_\theta(\boldsymbol{v}, \hat{\boldsymbol{z}_s}, \hat{\boldsymbol{z}_a}), \mathbf{I}_r\right).\end{aligned} \tag{10}$$

We then update the shape code by minimizing:

$$\begin{aligned}\mathcal{L}_s = &\left\|\hat{\mathcal{F}}_\theta(\hat{\boldsymbol{v}}, \boldsymbol{z}_s + \lambda_n \boldsymbol{z}_n, \hat{\boldsymbol{z}_a}) - \mathbf{I}_r\right\|_2 + \\ &\lambda_s D_{\text{CLIP}}\left(\hat{\mathcal{F}}_\theta(\hat{\boldsymbol{v}}, \boldsymbol{z}_s + \lambda_n \boldsymbol{z}_n, \hat{\boldsymbol{z}_a}), \mathbf{I}_r\right),\end{aligned}$$

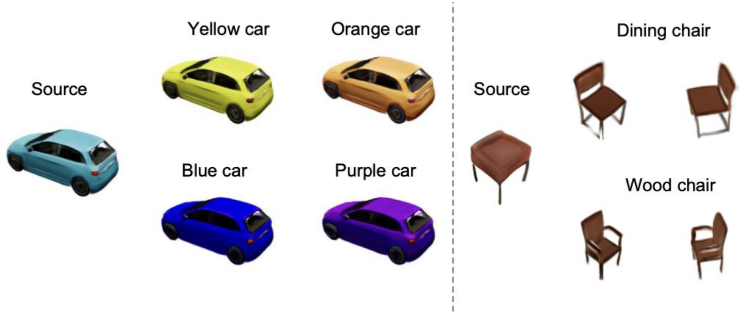
$$\begin{aligned}\mathcal{L}_a = &\left\|\hat{\mathcal{F}}_\theta(\hat{\boldsymbol{v}}, \hat{\boldsymbol{z}_s}, \boldsymbol{z}_a + \lambda_n \boldsymbol{z}_n) - \mathbf{I}_r\right\|_2 + \\ &\lambda_a D_{\text{CLIP}}\left(\hat{\mathcal{F}}_\theta(\hat{\boldsymbol{v}}, \hat{\boldsymbol{z}_s}, \boldsymbol{z}_a + \lambda_n \boldsymbol{z}_n), \mathbf{I}_r\right),\end{aligned}$$

e fixed, $\boldsymbol{z}_n$ is a random standard Gaussian
led in each step to improve the optimiza-
ıd λ_n linearly decays from 1 to 0 through

Experiments

Dataset

- CARLA(10k cars, 256x256)
- Photoshapes(150k chairs, 128x128)

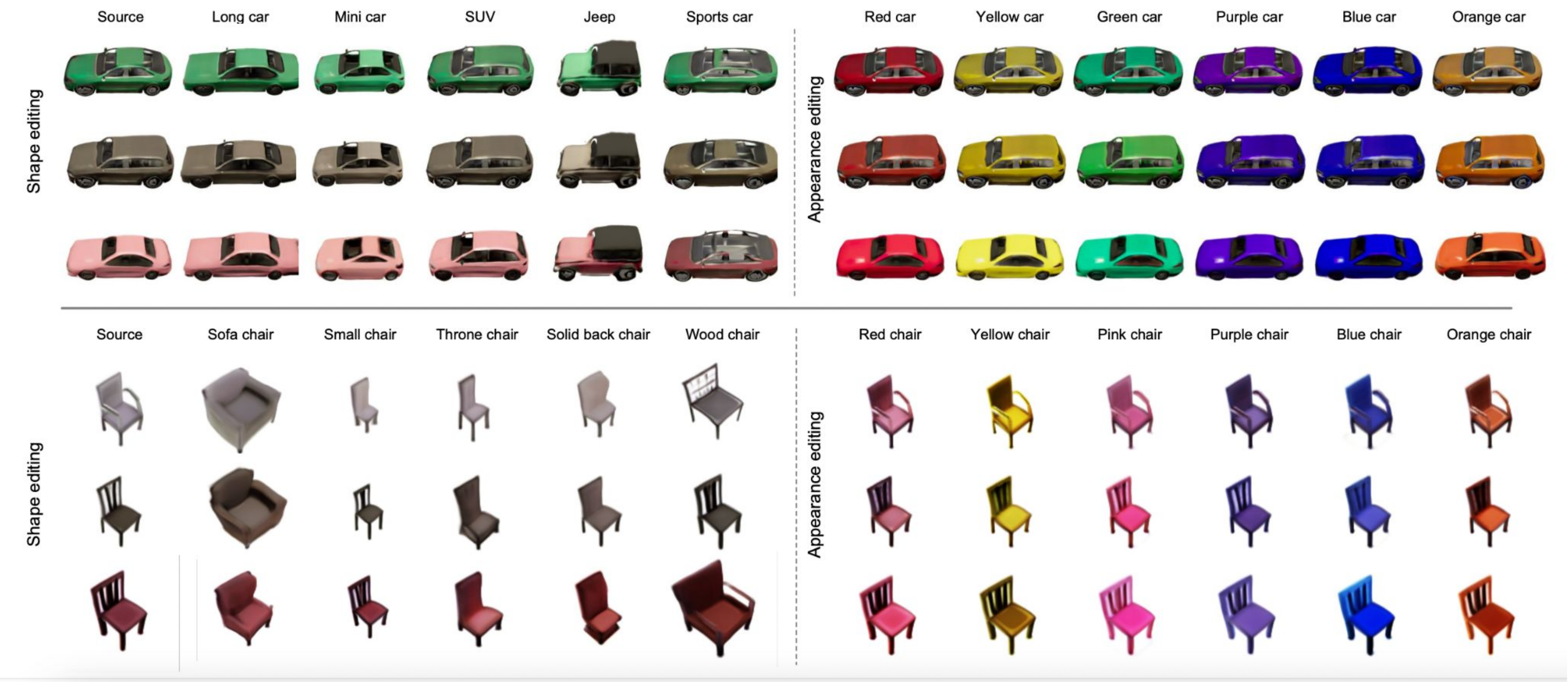


Inference time

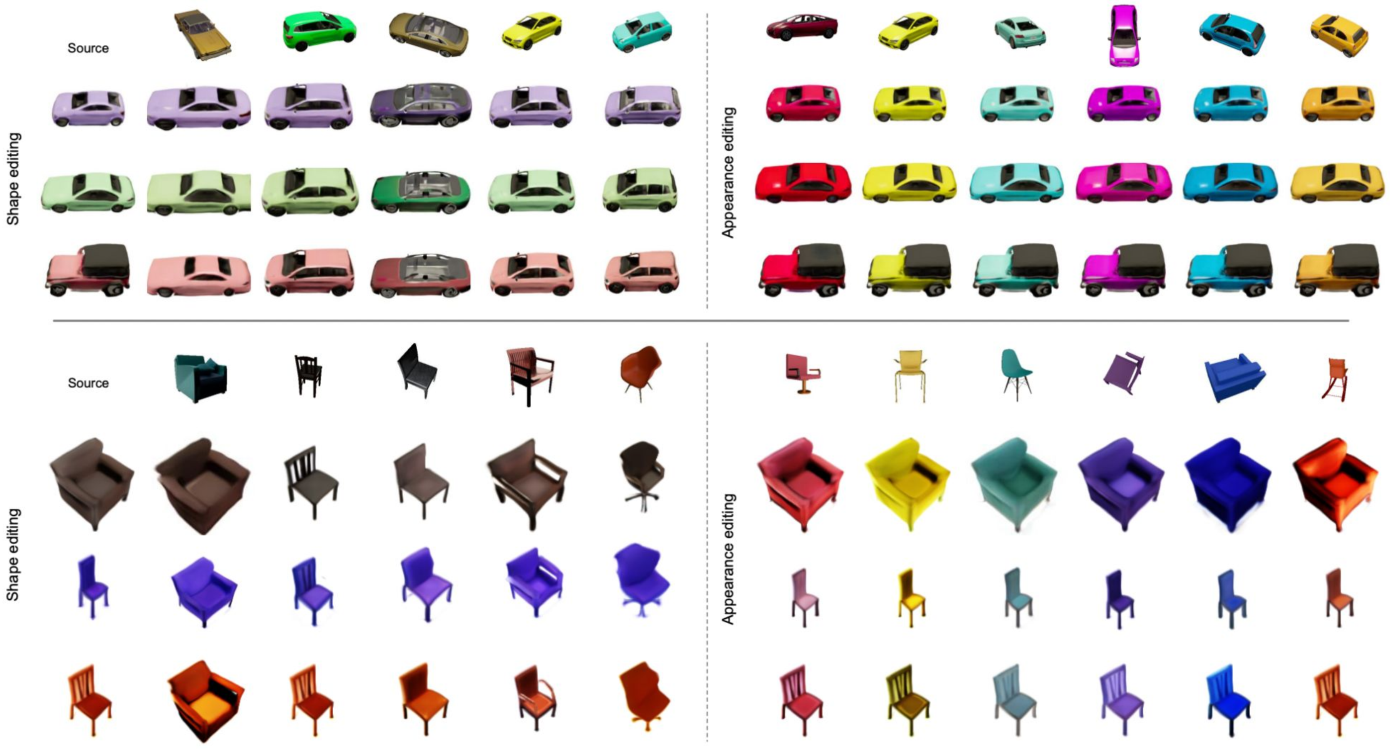
	Chairs		Cars	
	Shape	Appearance	Shape	appearance
EditNeRF	30.0	15.9	33.2	16.8
Ours	0.58	0.51	2.12	1.98

Table 1. **Compared to EditNeRF [21] on editing time averaged on 20 images.** We only include the inference/optimization time(s) and single-view rendered time(s) for chairs (128×128 pixels) and cars (256×256 pixels).

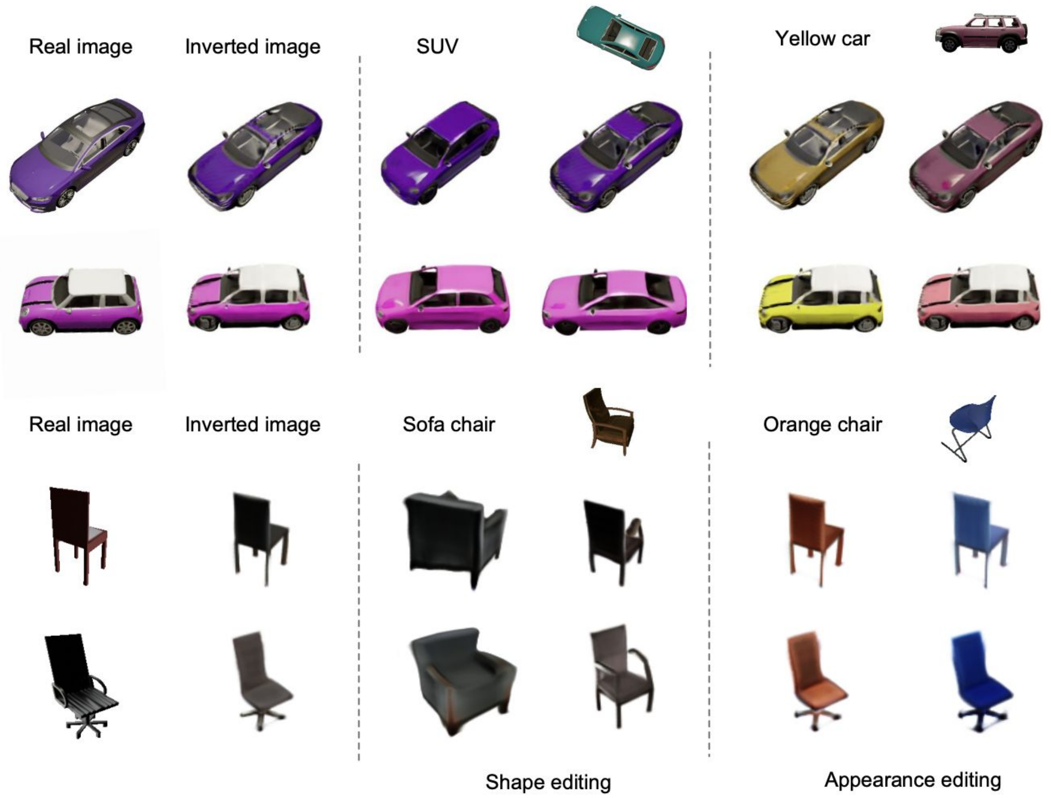
Results - text driven



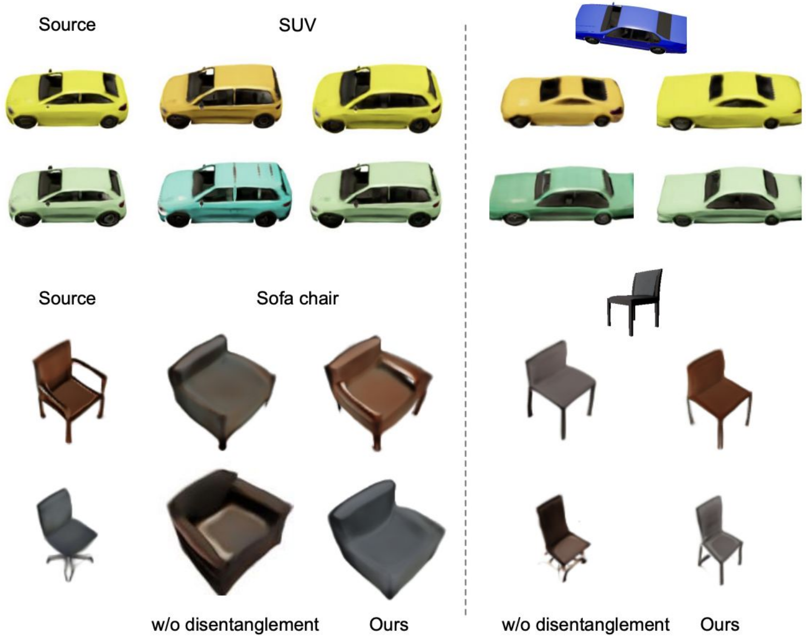
Results - exemplar driven



Results - Editing Real Images



Ablation



		Chairs			Cars		
		Before	After	Diff.	Before	After	Diff.
(a)	EditNeRF	36.8	40.2	3.4	102.8	118.7	15.9
	w/o disen.	52.5	54.3	1.8	69.2	69.9	0.7
	Ours	47.8	49.0	1.2	66.7	67.2	0.5
(b)	w/o disen.	52.5	53.2	0.7	69.2	71.1	1.9
	Ours	47.8	48.4	0.6	66.7	67.8	1.1

Table 2. **Fréchet inception distance (FID) for evaluating the image quality of reconstructed views before and after editing on: (a) color and (b) shape (lower value means better).** We use 2K images with various views drawn randomly from the latent space to calculate the FIDs for reconstructed images, and then perform various edits on these images to recalculate FIDs of edited results.

Ablation

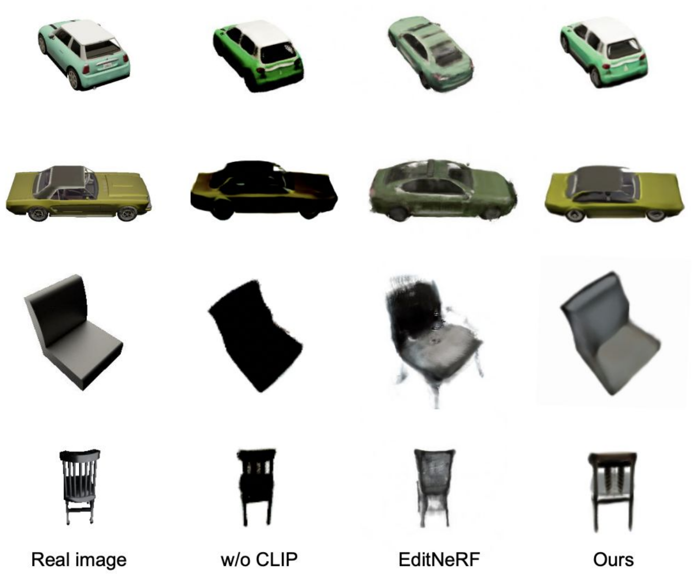


Figure 5. **Ablation study on our inversion method and comparison with EditNeRF.**

Limitations

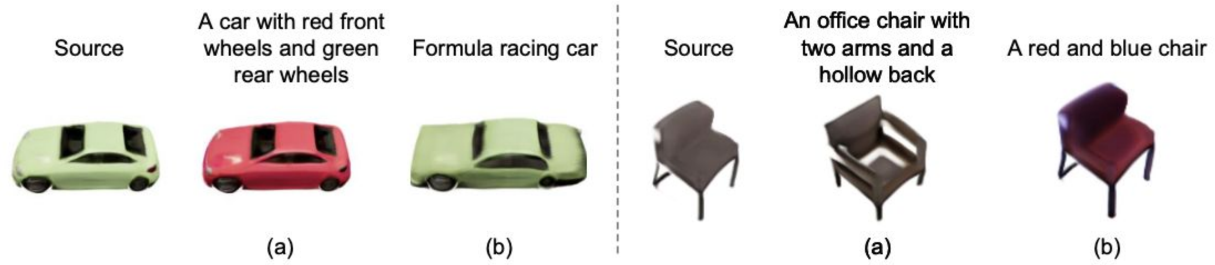


Figure 11. **Limitations.** Our method cannot handle fine-grained edits (a) and out-of-domain edits (b).