Pix2NeRF: Unsupervised Conditional π-GAN for Single Image to Neural Radiance Fields Translation

Supplementary Material

Method	128 × 128 FID ↓	KID ↓	IS ↑	64 × 64 PSNR ↑	SSIM ↑
Pix2NeRF unconditional	26.45	1.18	4.39	-	-
Pix2NeRF conditional	26.81	1.23	4.27	18.75	0.82

Table 4. Additional quantitative results on ShapeNet-SRN [4, 38].

Method	CelebA 64 × 64 FID ↓	KID ↓	IS ↑	ShapeNet-SRN 64 × 64 FID ↓	KID ↓	IS ↑	PSNR ↑	SSIM ↑
A	28.90	2.99	1.62	34.01	1.73	3.65	15.91	0.71
B	43.19	2.84	1.33	43.06	2.49	2.92	16.27	0.71
C	39.42	3.07	1.65	41.47	2.80	2.96	15.14	0.68
D	33.92	2.84	1.87	35.72	1.74	3.75	16.81	0.77
E	31.31	2.75	1.95	21.67	0.89	4.35	18.03	0.79
F	39.86	3.18	1.73	27.70	1.22	4.09	16.98	0.77
G	73.52	7.47	1.91	27.10	1.31	4.26	17.77	0.79
H	73.03	7.08	1.97	41.11	2.27	3.34	14.98	0.74
I	140.25	16.33	1.79	184.10	17.19	2.55	10.95	0.59
J	168.59	18.89	1.50	266.64	30.29	1.98	10.28	0.47
Full	**24.64**	**1.93**	**2.24**	**17.55**	**0.59**	**4.36**	**18.75**	**0.82**

Table 5. Quantitative results of ablation study on CelebA [19] and ShapeNet-SRN [4,38]. "Full" denotes Pix2NeRF conditional setup.

Table 6. Input view reconstruction (PSNR, SSIM) on a test set, and novel view synthesis (FID, KID×100, IS).

Method	PSNR↑	SSIM↑	FID↓	KID↓	IS↑
Pix2NeRF E + frozen π-GAN G	13.04	0.46	28.25	2.97	1.52
π-GAN optimization (200 iterations)	23.42	0.80	16.09	0.83	2.10
π-GAN optimization (700 iterations)	24.21	0.82	17.14	0.72	2.14
Pix2NeRF (feed-forward)	17.95	0.67	24.82	1.93	2.21
Pix2NeRF (200 iterations)	27.12	0.89	12.86	0.64	2.27
Pix2NeRF (1000 iterations)	**27.73**	**0.90**	**12.01**	**0.62**	**2.30**

A. Additional qualitative results

We demonstrate additional qualitative results achieved by Pix2NeRF on three datasets: CelebA [19], Shapenet-SRN chairs [4, 38], and CARLA [8] in Figures 6, 7, and 8 respectively.

B. Additional quantitative results

Table 4 provides additional quantitative results on ShapeNet-SRN [4, 38] with generative metrics computed on 128 × 128 resolution, and reconstruction metrics computed on 64 × 64 resolution. We do not report PSNR and SSIM for CelebA [19] as there is no ground truth novel views.

C. Additional ablation study

We provide quantitative results of each ablation study on CelebA [19] and Shapenet-SRN [4, 38] to further verify our design choices. As in the ablation study in our main paper, we report FID [15], KID [1] and IS [34] for CelebA [19], and additionally report PSNR and SSIM [42] on Shapenet-SRN [4,18]. We measure results after inference on resolution 64 × 64. We show quantitative ablation results in Table 5. Legend: **A** – naive GAN inversion; **B** – auto-encoder; **C** – no GAN inversion; **D** – no conditional adversarial objective; **E** – no warm-up; **F** – always warm-up; **G**, **H**, **I**, **J** – lower weights for reconstruction instead of warm-up, with $\lambda_{\text{recon}} = 1, 0.1, 0.01, 0.001$ respectively. Note that since the encoder output is not enforced to strictly follow p_z, naive GAN inversion (stage 1 in [31]) failed completely due to bad initialization. We therefore use a "warmed-up" version of the generator trained for 300k iterations.

D. Input reconstruction and hybrid optimization

We ran extra ablations and summarized our model performance by providing both input reconstruction (cols 2,3) and novel view synthesis (cols 4,5,6) results in Tab. 6 (row 4). We show π-GAN latent optimization on an input image for 700 iterations, as recommended by its authors in row 3. Note that it requires time-consuming per-instance optimization due to the NeRF's rendering mechanism. Additionally, we use the Pix2NeRF encoder's output as a starting point and perform latent optimization with a frozen Pix2NeRF generator for only 200 iterations, shown in row 5. A qualitative comparison is shown in Fig. 5. Note that our model does not overfit the input view even with 1000 iterations of input view optimization (row 6), while π-GAN shows strong artifacts and requires a search for the optimal number of iterations.

E. Necessity of generator distilling

We trained the encoder with a pretrained frozen π-GAN generator using all the losses. As can be seen from the results in Tab. 6 Row 1, the model struggles to capture details accurately without fine-tuning the generator jointly.

F. Linear interpolation

We interpolate novel views between two different input images by predicting their corresponding latent codes and poses, then applying linear interpolation to get the intermediate codes and poses. We show the results interpolating five images in Figure 9.

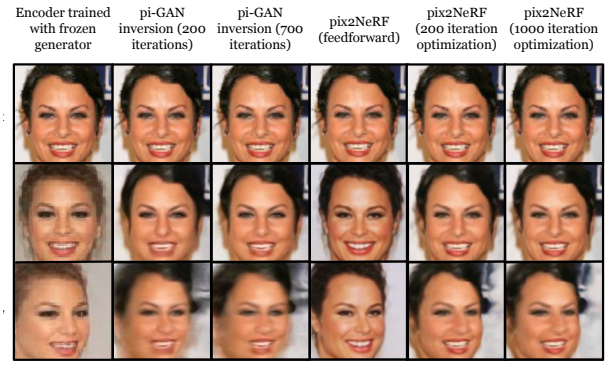

Figure 5. Qualitative comparison on CelebA. Top – input, middle – reconstruction, bottom – novel view synthesis.

G. Limitations and failure cases

Despite training on images without pose or 3D supervision, Pix2NeRF can reconstruct objects from a single image and achieve decent quality. However, the methodology of using an encoder to encode an entire image into a single latent code is quite challenging, especially when the dataset is noisy, such as CelebA [19]. Pix2NeRF cannot always capture fine details accurately. We observe failure cases when the input is out-of-distribution relative to that of the training set p_{real}, as shown in Figure 10. It might be possible to improve these hard cases by introducing pixel-wise features instead of (or, in addition to) the global latent code, as done in PixelNeRF [46] and GRF [40].

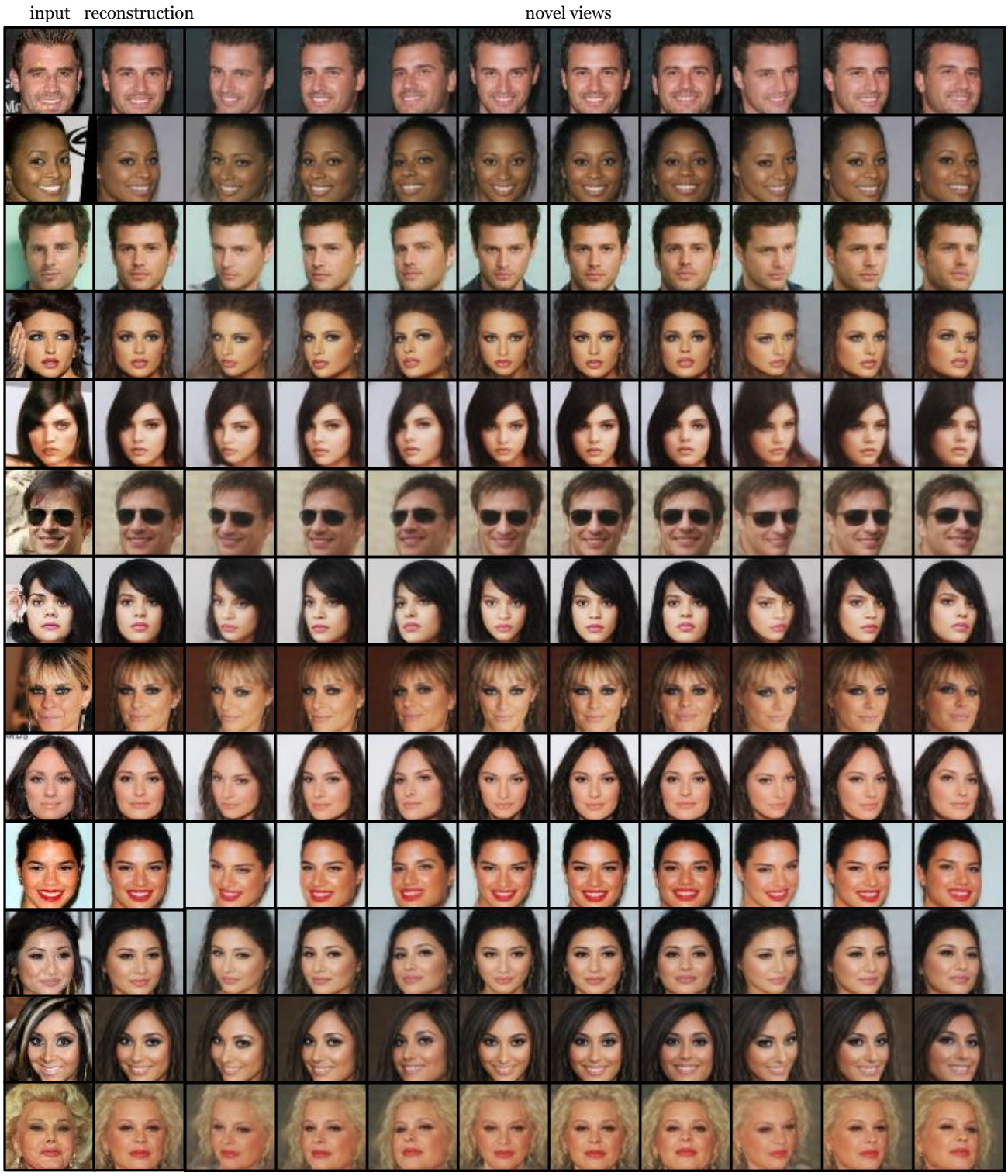

Figure 6. Further reconstructions and novel views on CelebA [19].

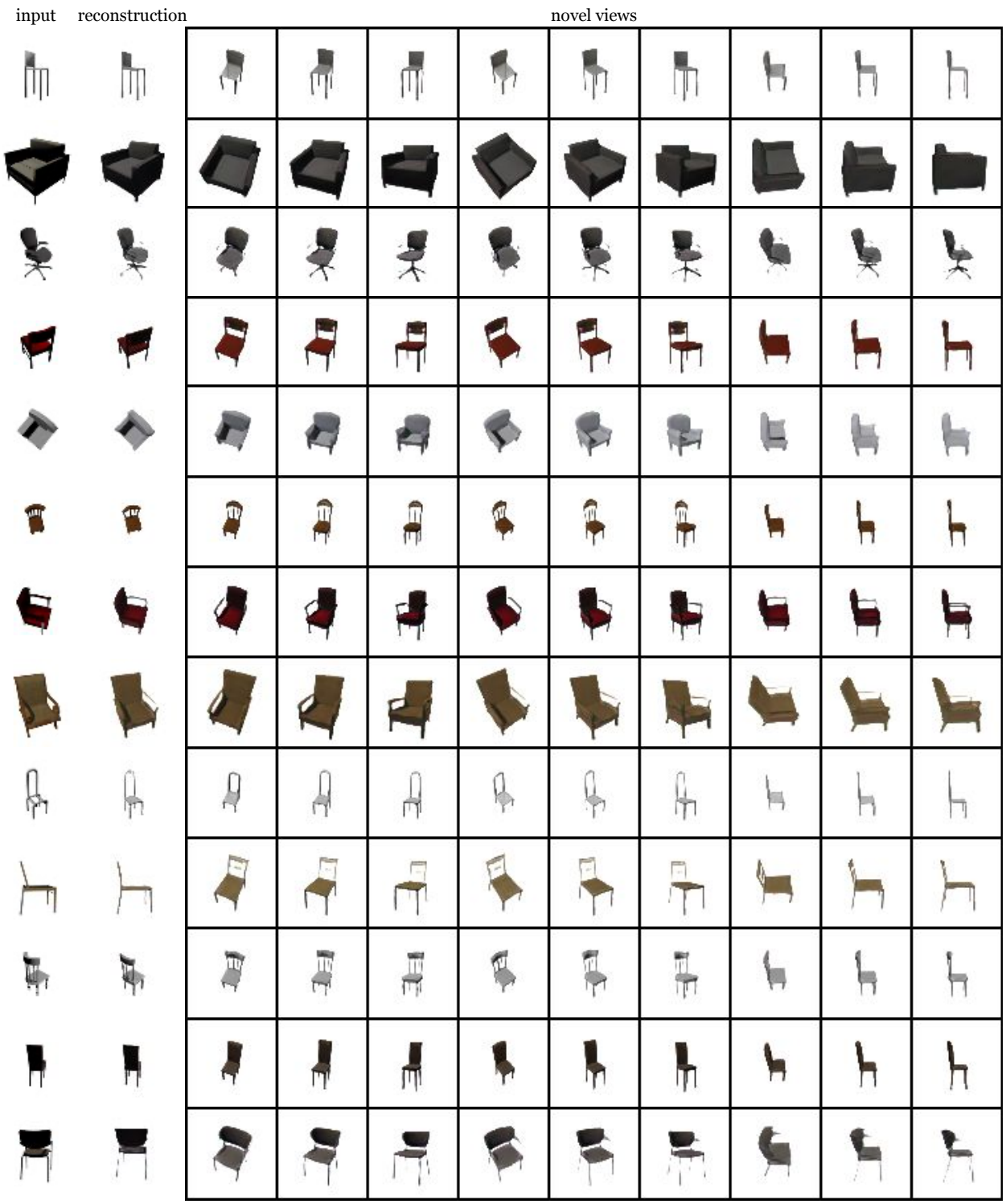

Figure 7. Further reconstructions and novel views on ShapeNet-SRN [4, 38].

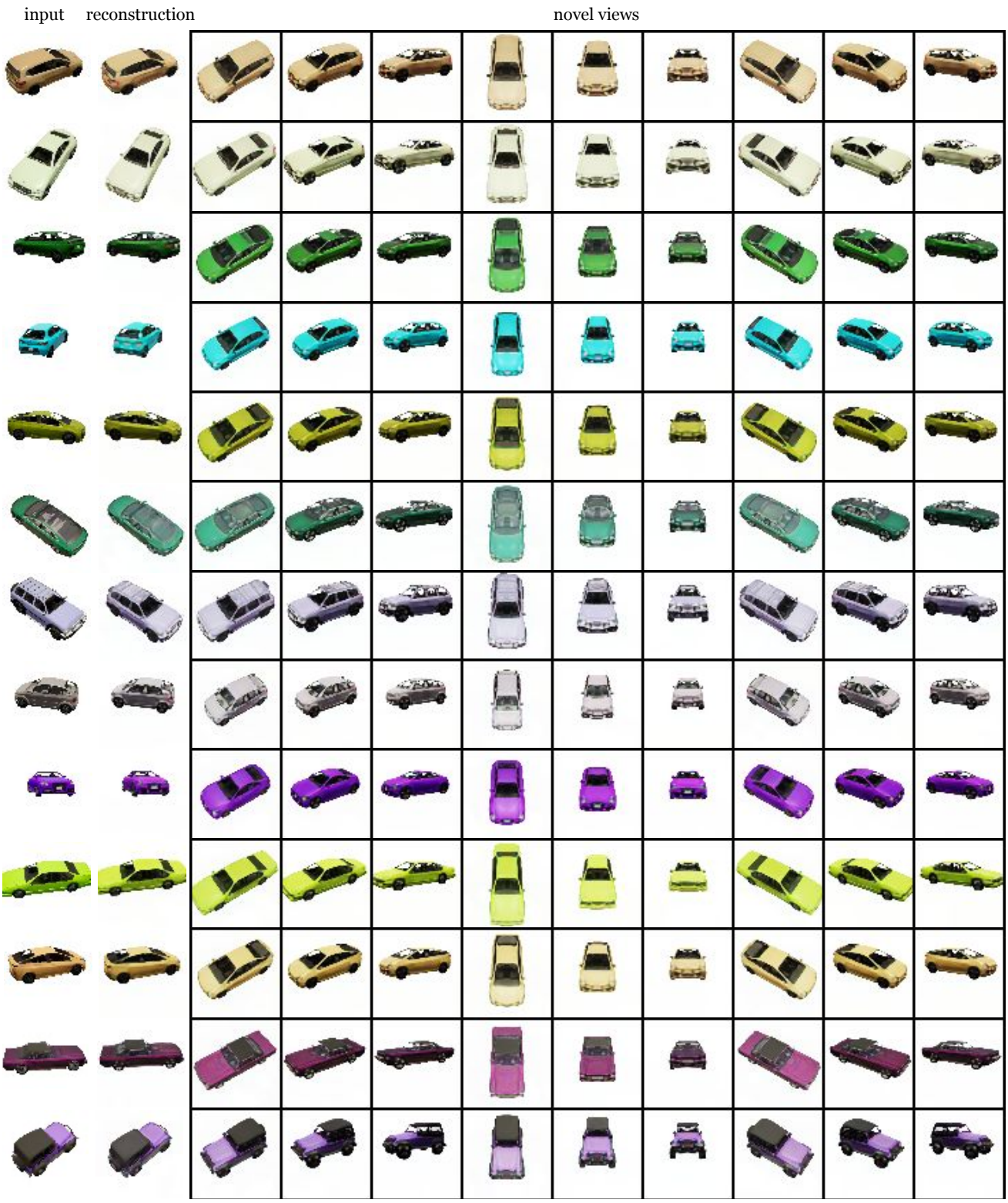

Figure 8. Further reconstructions and novel views on CARLA [8].

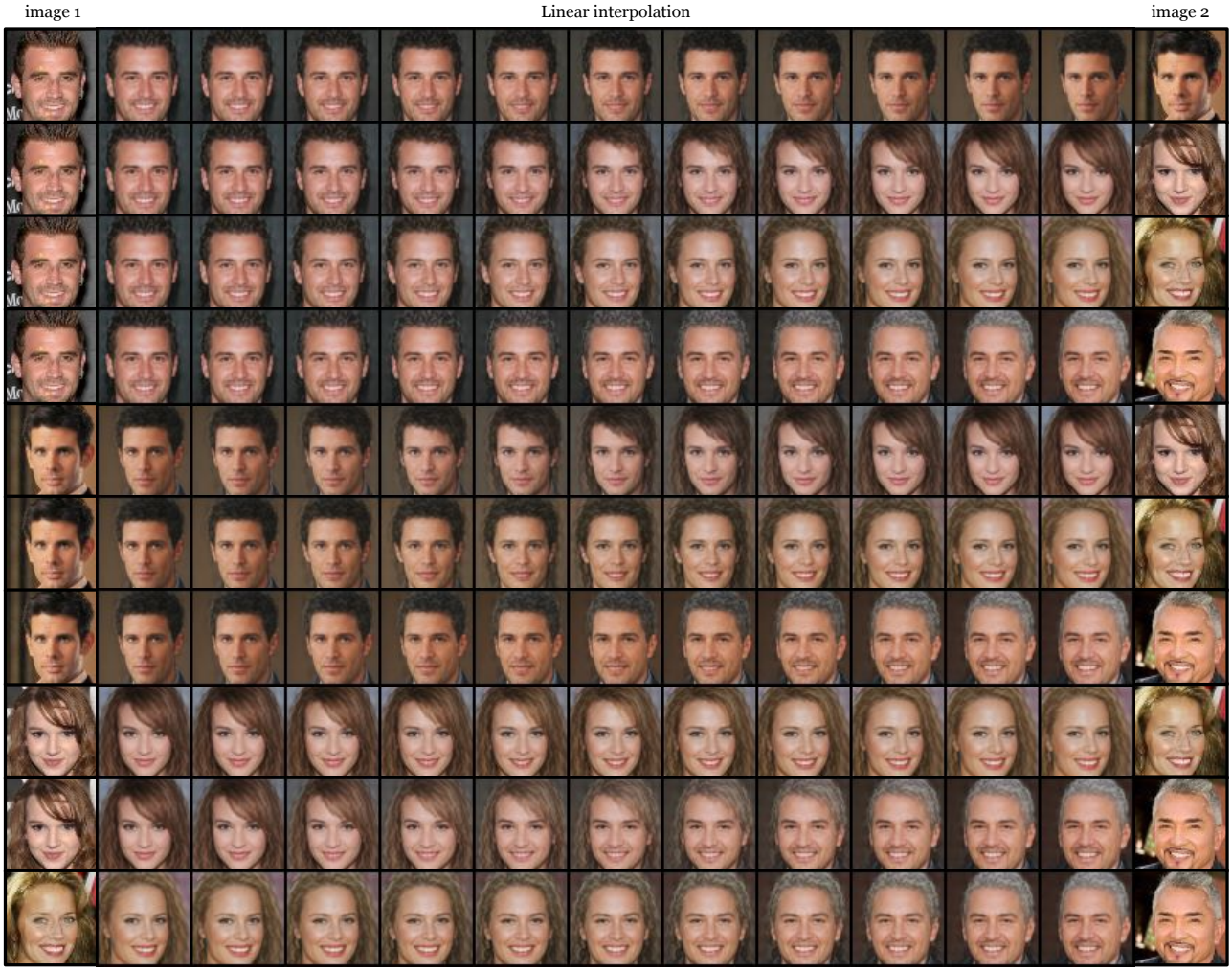

Figure 9. Linear interpolation on CelebA [19].

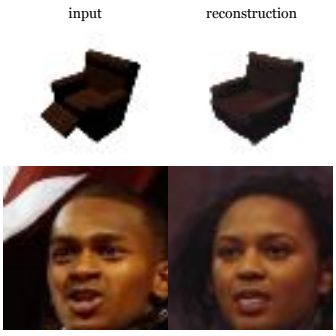

Figure 10. Failure cases on CelebA [19] and ShapeNet-SRN [4, 38].