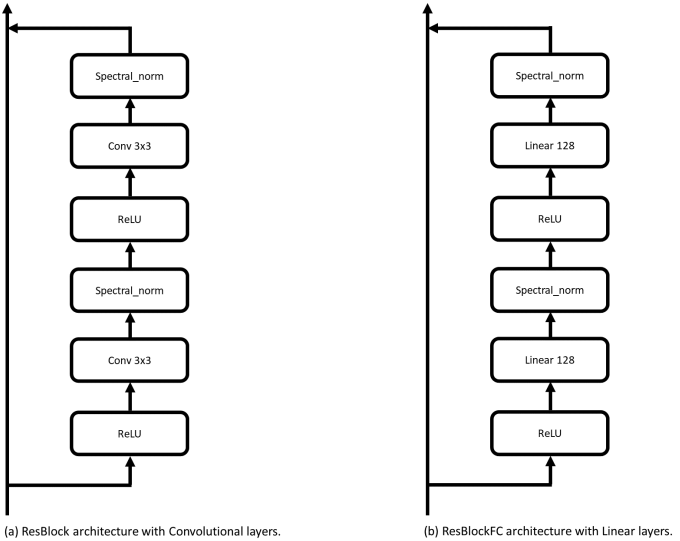
Supplementary Material: Class-Continuous Conditional Generative Neural Radiance Field

Jiwook Kim
tom919@cau.ac.kr
Minhyeok Lee*
mlee@cau.ac.kr

School of Electrical & Electronics Engineering
Chung-Ang University
Seoul 06974, Republic of Korea

1 Model Details

In this section, we discuss our detailed network architectures with residual modules. We adopt the residual modules to the discriminator, decoder, and neural renderer as Figure 1, 2, 3.



(a) ResBlock architecture with Convolutional layers. (b) ResBlockFC architecture with Linear layers.

Figure 1: **Architectures of ResBlocks used in our model.** We employ two types of residual modules, the ResBlock and ResBlockFC. While the ResBlock is comprised of convolution layers, the ResBlockFC is based on linear layers. Therefore, we utilize the ResBlock for the discriminator and the neural renderer, which require spatial information, whereas the ResBlockFC is utilized for the decoder, which maps a 3D coordinate, a viewing direction, and conditional encodings to a feature space and a volume density with linear layers.

*Corresponding author.

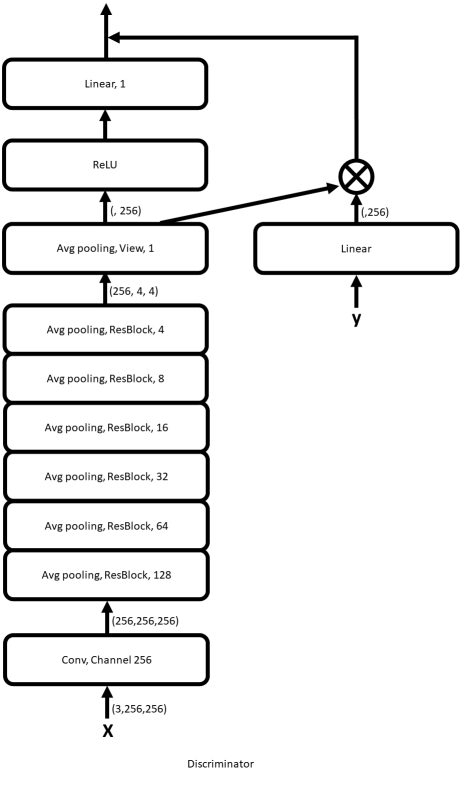


Figure 2: **Architecture of the discriminator.** The **x** and **y** indicate an input image of the discriminator and a conditional label, respectively. Each **x** and **y** are embedded with the convolutional layer and the linear layer. The embedded image is encoded by passing several average pooling layers and ResBlocks. The encoded image is projected to the embedded conditional label. By the projection, we can fuse information of the image and conditional label.

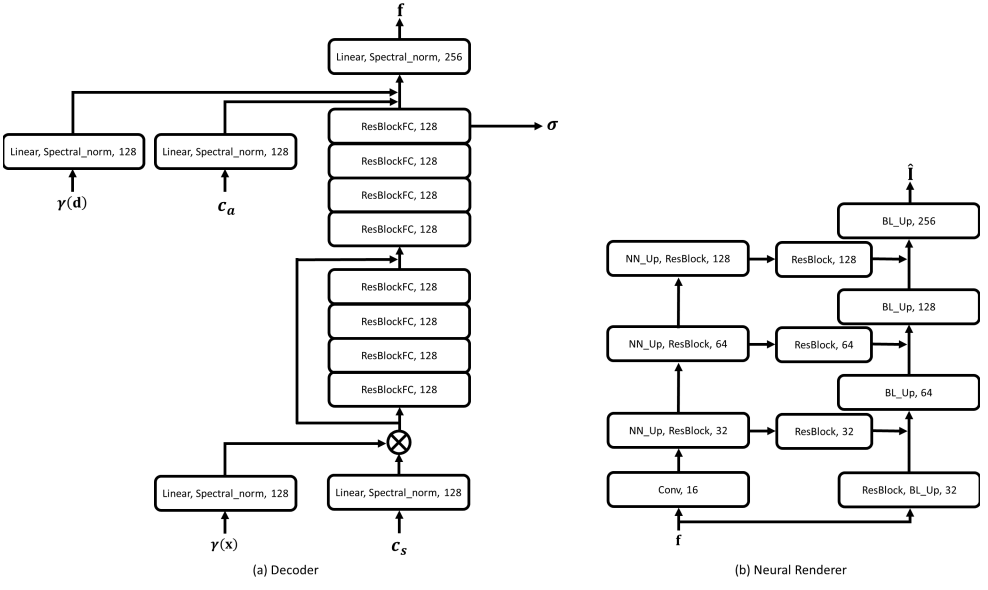


Figure 3: **Architectures of the decoder and neural renderer.** The (a) and (b) show the architecture of the decoder and neural renderer, respectively. In (a), the shape conditional encoding $\mathbf{c_s}$ and the positional encoded 3D point $\gamma(\mathbf{x})$ are embedded and multiplied. We use 8 blocks of the ResBlockFC and one skip-connnection. After passing the blocks, the volume density σ is estimated as the output of the final ResBlockFC. By adding the appearance conditional encoding $\mathbf{c_a}$ and the positional encoded viewing direction $\gamma(\mathbf{d})$ to the volume density and passing the linear layer, the decoder outputs the feature $\mathbf{f}$. In (b), BL_Up and NN_Up represent a bilinear upsampling and a nearest neighbor upsampling, respectively. The neural renderer takes the feature $\mathbf{f}$ as an input and outputs a synthesized image $\hat{\mathbf{I}}$.

2 Controllable Features in 3D Object Generation

In this section, we report additional 3D-aware generations with controlling the features as well as a negative result as Figure 4, 5.

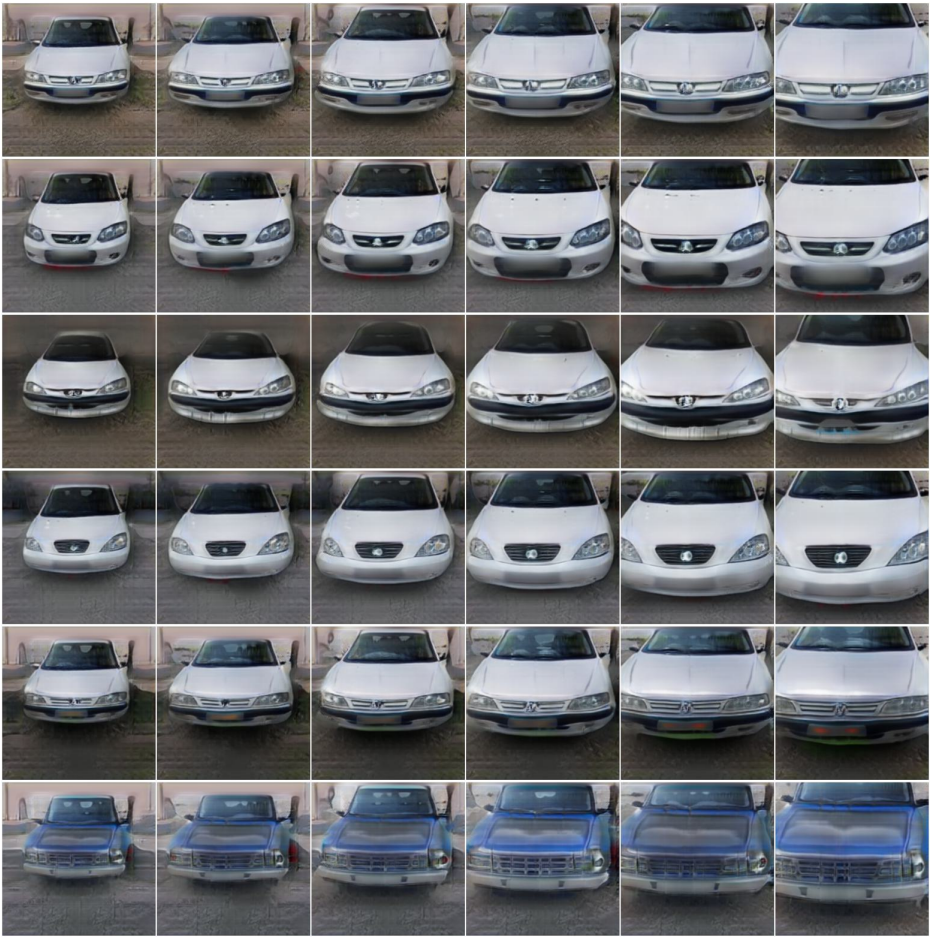


Figure 4: **Visualization of manipulating the scaling.** Each row and column indicate a single object with the same conditional class and the same scale value, respectively. Our model generate a object with various scales, including out-of-distribution images.

Figure 5: **Negative result.** Each column indicates different rotation angles. Near to the edge of columns represents more rotation. While we trained with 60° rotation angle, we produce the results with 120° rotation angle, corresponding to out of distribution. The excessive out of distribution occurs negative results as the edge of column images.

3 Continuously Controllable Features in 3D Object Generation

We report additional interpolation and extrapolation results on CelebA with manipulations of different labels as Figure 6, 7, 8, 9, 10, 11.

Figure 6: **Interpolation and extrapolation results with the glass condition.** Each column indicates generated images with the corresponding label values between zero and three.

Figure 7: **Interpolation and extrapolation results with the aging condition.** Each column indicates generated images with the corresponding label values between zero and three.

Figure 8: **Interpolation and extrapolation results with the make-up condition.** Each column indicates generated images with the corresponding label values between zero and three.

Figure 9: **Interpolation and extrapolation results with the rosy cheek condition.** Each column indicates generated images with the corresponding label values between zero and three.

Figure 10: **Interpolation and extrapolation results with the bald condition.** Each column indicates generated images with the corresponding label values between zero and three.

Figure 11: **Interpolation and extrapolation results with the mustache condition.** Each column indicates generated images with the corresponding label values between zero and three.

4 Qualitative Evaluation of Residual Modules

We compare C3G-NeRFs with residual modules and conditional GIRAFFE with a plain network in terms of generated image quality. This experiment demonstrates the effect of adopting residual modules in C3G-NeRF architecture. As shown in Figure 12, C3G-NeRF generates high-quality images with fine details after a small number of iterations. In contrast, the conditional GIRAFFE using plain networks shows poor generation quality as seen in Figure 13 with more (five times larger) iterations. Therefore, we can confirm that it is necessary to use the proposed residual modules in the NeRF structures in order to generate high-quality conditional images.

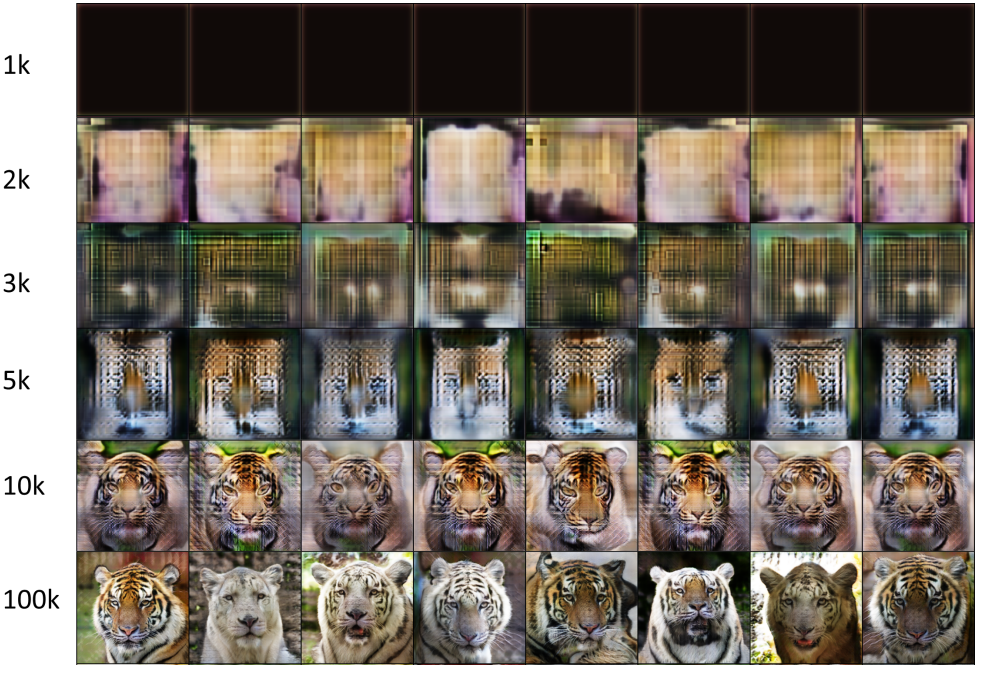


Figure 12: **Training results in terms of iteration.** By employing residual modules, our model produces diverse and high-quality images with fine details within a small number of iterations.

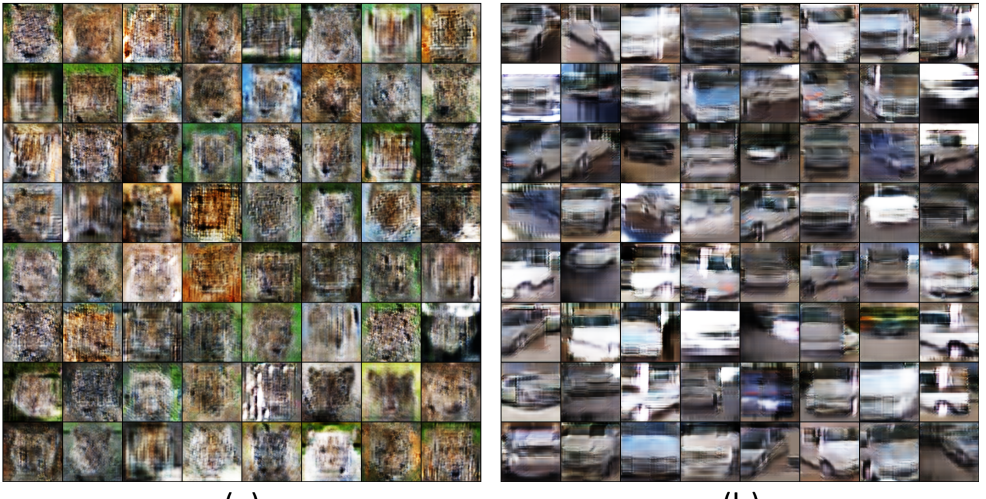


Figure 13: **Generated objects by the conditional GIRAFFE with plain networks trained with 500,000 iterations.** (a) and (b) show the synthesized images with plain networks-based conditional GIRAFFE trained in AFHQ and Cars, respectively. Despite enough iterations (five times larger compared to our model), the generator synthesizes low-quality images.

Figure 14: **Random synthetic samples on AFHQ with a 256^2 resolution.**

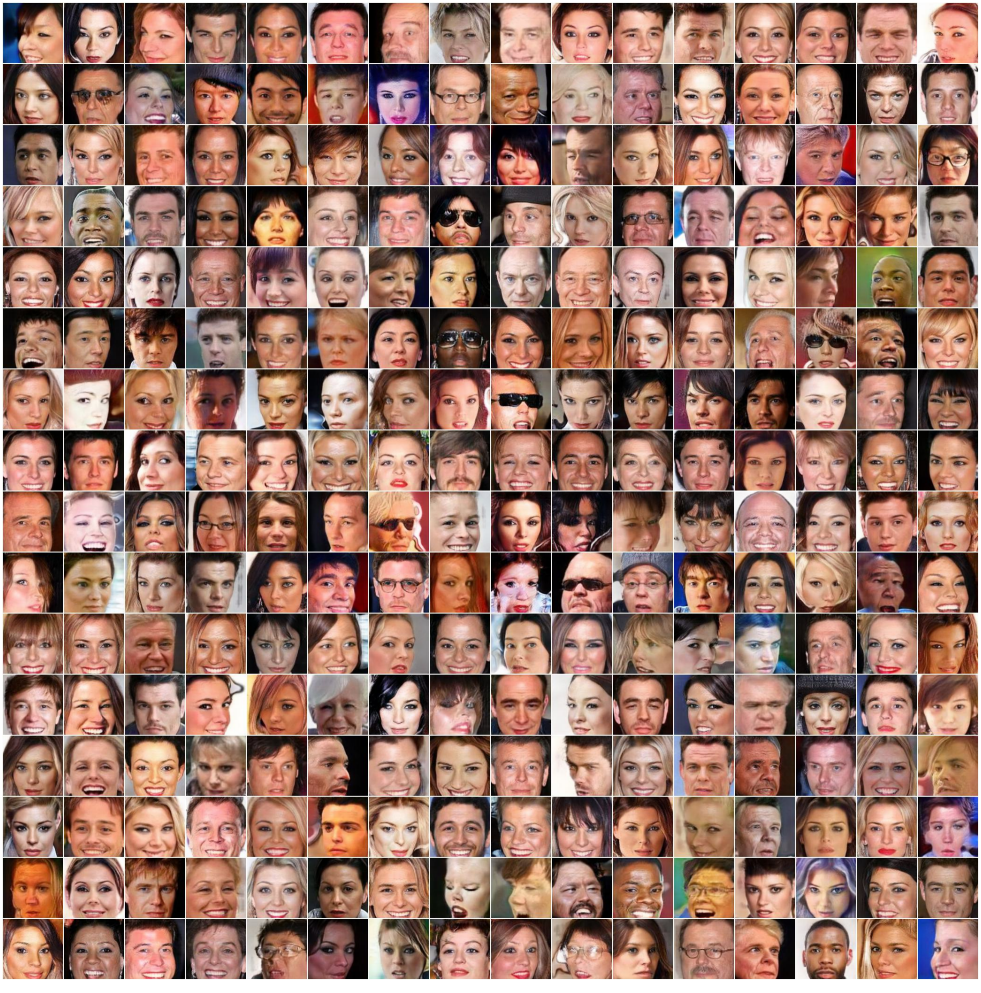


Figure 15: **Random synthetic samples on CelebA with a 128^2 resolution.**

Figure 16: **Random synthetic samples on Cars with a 256^2 resolution.**

5 Acknowledgement

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00251528).