FDNeRF: Semantics-Driven Face Reconstruction, Prompt Editing and Relighting with Diffusion Models

Anonymous Author(s)
Affiliation
Address
`email`

Please also watch the supplementary video for visualization of results at various viewpoints.

1 Implementation Details

Our reconstruction loss and identity loss are applied to the ground truth image camera poses. Due to our limited GPU memory, we only render one side view to calculate the diffusion loss and illumination loss at each iteration. The camera rotation angles θ and ϕ are randomly sampled from $[\frac{\pi}{2}-\frac{\pi}{12}, \frac{\pi}{2}+\frac{\pi}{12}]$ and $[\frac{\pi}{2}-\frac{\pi}{12}, \frac{\pi}{2}+\frac{\pi}{12}]$, where θ and ϕ are the angles of spherical coordinate. We set the optimization iterations for our editing to 500, which takes approximately 10 minutes on a 3090 GPU. We set the weighting $\mathcal{L}_D$, $\mathcal{L}_{ID}$, L_R to be 0.2, 0.2 and 2×10^{-5} for most editing cases, which can be finetuned for each editing.

Dataset and Generative Bias We utilize trained checkpoints of EG3D on FFHQ [3], AFHQv2 [2] and ShapeNet [1] for the data domain of face, cat and car respectively. For some editing and generation, we notice the existence of biases in generated results caused by the bias of the training dataset, especially for race, gender, etc.

2 More Results

As a supplement to Figure 1 and Figure 3 in our main paper, fig. 1 shows more results of our FDNeRF. All three figures illustrate that given a single face image, our FDNeRF can perform semantics-driven NeRF editing on various features, such as expressions, emotions, glasses, hairstyles, races, genders, ages, makeup, beard, mustache, and goatee. Notably, both individual and joint editing of these features can be achieved in high fidelity.

Text-conditioned Generation Comparison Our attempts to compare with Dreamfusion[4] (Implemented on a StableDiffusion) failed since it cannot generate faithful models for human heads. This problem might be caused by various reasons. First, the ambiguity of text prompts (human identity, rendering directions) might result in an inconsistent denoising behavior at each iteration. Also, the randomized and unconstrained NeRF optimization process might collapse. On the contrary, our utilization of latent code and trained 3D generative models ensures successful and high-quality generation.

3 Ablation Study on Editing Prompts

We investigate the influence of the input text prompt in this section. In fig. 1, and Figure 1 and 3 in our main paper, we mainly show NeRF editing results when the input text prompt is a short sentence or a full description of the expected NeRF, such as “A woman wearing dark red lipstick”. While this helps a valid generation that avoids ambiguity in semantics, our FDNeRF can also take in a text prompt that only expresses the difference between the input face and the output face, such as “Dark red lipstick”, if the input face has no lipstick at all. Here, fig. 2 shows the ablation results. The right

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

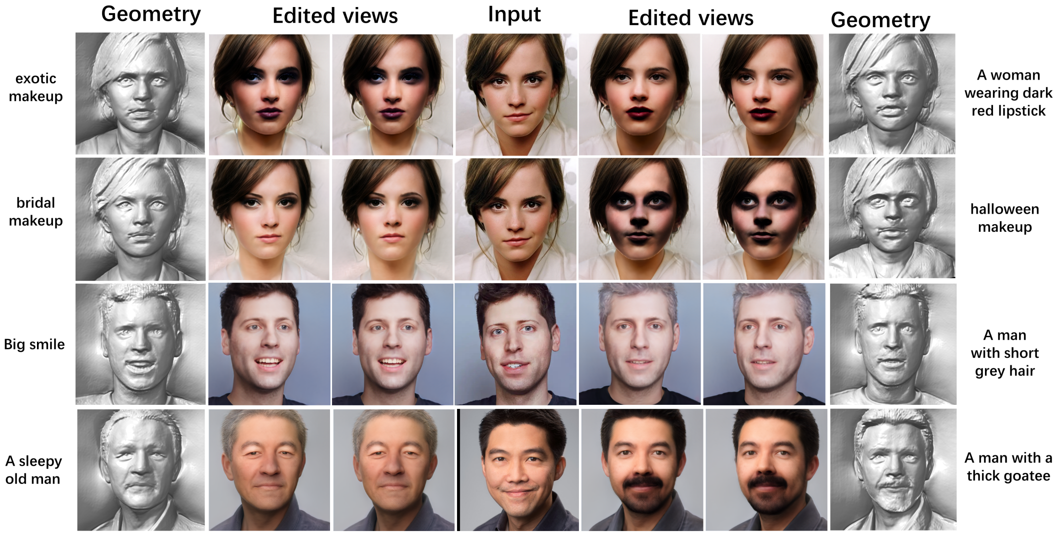

Figure 1: **More FDNeRF results.** The middle column shows the input images, while the left and right halves of the images in the other two columns show the results of text prompts editing on the left and right, respectively. Additionally, the first and last columns showcase the corresponding geometries.

two columns show NeRF editing results when input text prompts are full descriptions of the expected editing results, while the left two columns are results with text prompts only describing what should be different.

In most cases, our FDNeRF can generate similar editing results given either type of text prompt with subtle or even no tuning of weights of losses λ_{ID}, λ_R, and λ_D, as shown in fig. 2. However, sometimes ambiguity in semantics interferes editing when the input text prompt is not a full description of the expected output. As shown in fig. 3, even though our FDNeRF can generate an old Elon Musk given "An old man" as the input text prompt, it fails when the input becomes just "Old". We believe this is because "Old" has various meanings depending on its context. Therefore, a single "Old" without any context results in useless guidance from our diffusion model, thus producing poor editing results. "Elderly" and "senior" are two synonyms of "Old" in this context. As shown in fig. 3. "senior" also fails due to its semantic ambiguity with no context, while "elderly" succeeds because of its specific meaning.

Thus, users of our FDNeRF are advised to give a better and complete text prompt to avoid semantics ambiguity.

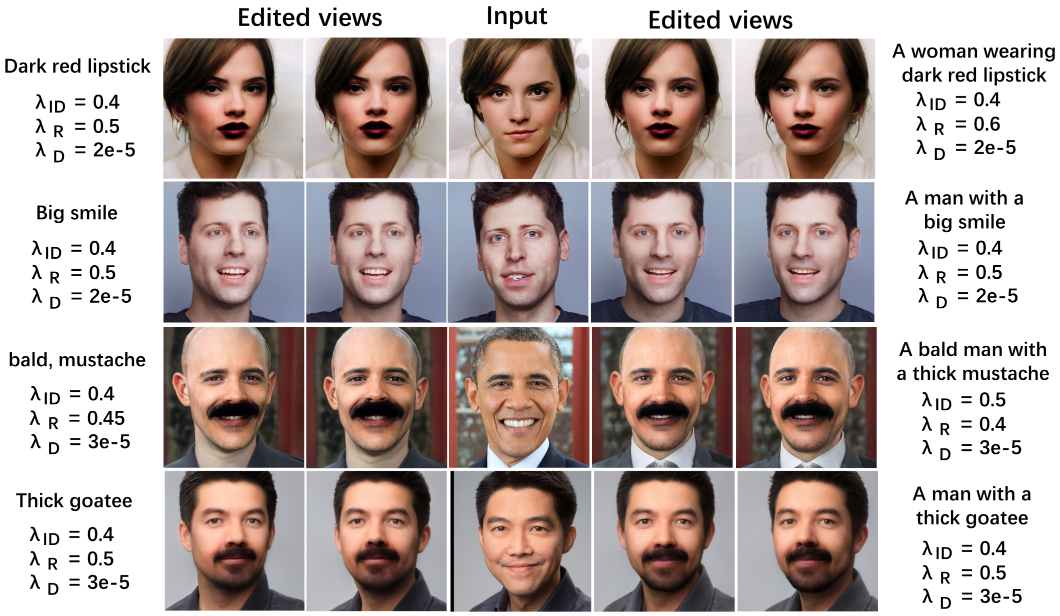

Figure 2: **Ablation study on the text prompt.** On the right are results generated with full text prompts. The left ones are produced using simlfied prompts. Given text prompt describing the wanted change, FDNeRF generates results with desired editing.

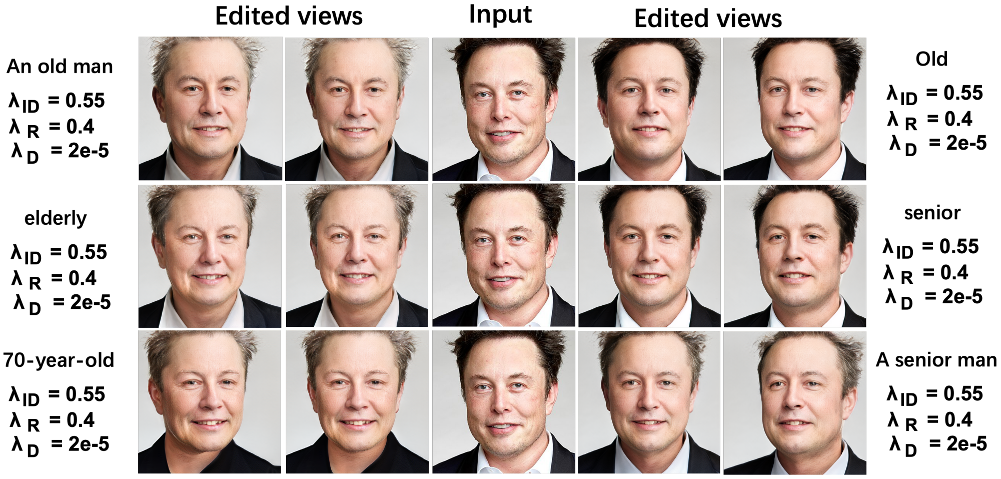

Figure 3: **Ablation study on synonyms.** Here are results generated with various Synonyms, which could affect the results. The contest is important for synonyms to eliminate ambiguity, as the case of "senior" and "A senior man".

References

[1] Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F., 2015. Shapenet: An information-rich 3d model repository. `arXiv:1512.03012`.

[2] Choi, Y., Uh, Y., Yoo, J., Ha, J.W., 2020. Stargan v2: Diverse image synthesis for multiple domains, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

[3] Karras, T., Laine, S., Aila, T., 2018. Flickr faces hq (ffhq) 70k from stylegan. CoRR URL: `https://github.com/NVlabs/ffhq-dataset/blob/93955b7cd435b7b1c724f8ca6a0e0c391300fe83/README.md`.

[4] Poole, B., Jain, A., Barron, J.T., Mildenhall, B., 2022. Dreamfusion: Text-to-3d using 2d diffusion. arXiv .