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

Hao Zhang * **Yanbo Xu** * **Tianyuan Dai** * **Yu-Wing Tai** **Chi-Keung Tang**
Hong Kong University of Science and Technology
{hzhangcc, yxubu, tdaiaa}@connect.ust.hk
yuwing@gmail.com
cktang@cs.ust.hk

Abstract

The ability to create high-quality 3D faces from a single image has become increasingly important with wide applications in video conferencing, AR/VR, and advanced video editing in movie industries. In this paper, we propose Face Diffusion NeRF (FDNeRF), a new generative method to reconstruct high-quality Face NeRFs from single images, complete with semantic editing and relighting capabilities. FDNeRF utilizes high-resolution 3D GAN inversion and expertly trained 2D latent-diffusion model, allowing users to manipulate and construct Face NeRFs in zero-shot learning without the need for explicit 3D data. With carefully designed illumination and identity preserving loss, as well as multi-modal pre-training, FDNeRF offers users unparalleled control over the editing process enabling them to create and edit face NeRFs using just single-view images, text prompts, and explicit target lighting. The advanced features of FDNeRF have been designed to produce more impressive results than existing 2D editing approaches that rely on 2D segmentation maps for editable attributes. Experiments show that our FDNeRF achieves exceptionally realistic results and unprecedented flexibility in editing compared with state-of-the-art 3D face reconstruction and editing methods. Our code will be available at https://github.com/BillyXYB/FDNeRF.

1 Introduction

Rich and versatile 3D contents are in high demand in entertainment industries such as movie making, computer gaming and emerging applications such as Metaverse, which also have high potential in robotics learning as well, where high-quality synthetic data in large amounts are required to enhance generalizability. 3D generative methods such as EG3D [6] can generate high-fidelity NeRF from a single image, but the reconstructed NeRF cannot be easily controlled or edited. Some methods [29, 51] use pixel-wise segmentation maps or user-supplied sketches to guide 3D editing. But these methods are hard to scale up due to the demanding editing requirement and limited editable attributes. Language is regarded as one of the most suitable candidates to provide control signals for 3D editing, especially given the current success of 2D semantics-driven editing [24, 38], where image and language domains are bridged by CLIP, i.e., Contrastive Language-Image Pre-training [41]. However, there is still substantial room for improvement in 3D generative and editable models, in terms of usability, edit-ability, and results quality.

Neural Radiance Field (NeRF), a new paradigm in 3D scene representation, embeds a 3D scene in a compact fully-connected neural network and achieves realistic rendering of novel views. NeRF is optimized to approximate a continuous and thus differentiable scene representation function, which has quickly become a dominant approach for many relevant downstream tasks including

*These authors contributed equally to this work.

Preprint. Under review.

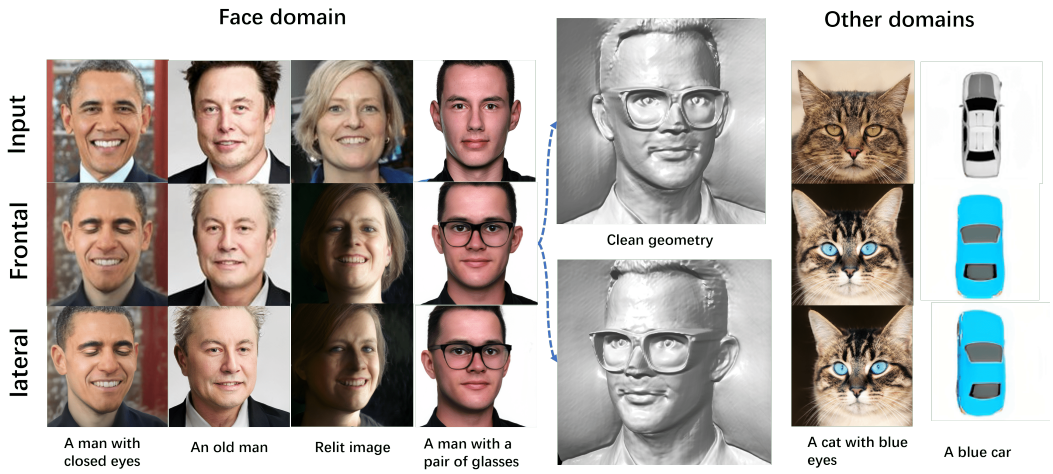

Figure 1: **FDNeRF results.** Given a face image, FDNeRF can reconstruct, edit, and relit a photo-realistic 3D face NeRF using a text prompt and a target light. Our method is not limited to human faces, and can be applied to other domains as well. Further results and videos are available in the supplemental materials.

novel scene composition [25, 34, 37, 58, 62], surface reconstruction [27, 54, 60], and articulated 3D shape reconstruction [8, 19, 39, 56, 59, 63] due to its differentiability, multi-view consistency, compactness, and fast inference. As NeRF is trained on multi-view images, relighting the NeRF depends on the relighting of the multi-view 2D images. However, the current 2D relighting method [4, 18, 30, 12] can't guarantee the view consistency of the relit multi-view images, which is crucial for real applications.

Significant attempts have been made in recent years to develop semantic-driven NeRF editing models. Pre-trained 2D language-image models are usually leveraged to enable multi-modality in 3D. CLIP-NeRF [53] introduces a disentangled conditional NeRF architecture, where the disentangled latent representations can be bridged by two mappers trained with a CLIP-based matching loss to the CLIP embedding, based on which the latent codes are updated for editing based on the input text prompt. Outperforming CLIP-based methods, DreamFusion [40] later adopts Imagen [44], a pre-trained text-to-image diffusion model, as a prior to guide the optimization of parameters of a randomly initialized NeRF through a novel SDS loss. Magic3D [28] further improves DreamFusion by introducing a coarse-to-fine optimization scheme with diffusion priors at different resolutions. However, these methods have two main limitations: lack of photorealism and long inference time (reportedly taking 1.5 hours for DreamFusion, and 40 minutes for Magic3D).

We believe the main reason underlying the less realistic results and slow inference of DreamFusion [40] and its variants lies in the random and thus unrestricted initialization of NeRF. Leveraging 3D generative models such as EG3D [6], whose generation is restricted by the discriminator during training for constrained NeRF generation, may offer a feasible solution. Although combining 2D generative models with CLIP [41] for realistic image generation and editing has been a widely adopted approach, no significant attempts have been made in 3D context to our knowledge. Thus, in this paper, we present FDNeRF, which is to our knowledge the first work to enable semantic-driven NeRF editing in 3D with high photorealism and versatile relighting from a single image, given a text prompt and target light. FDNeRF freezes the weights of the EG3D network which can reconstruct photorealistic NeRF from a single image. To enable semantic-driven editing, we adopt stable diffusion [43] to guide the optimization through SDS loss introduced in DreamFusion [40]. Based on SDS loss, identity loss, and feature loss, the latent vector in EG3D's latent space can be updated. Moreover, the proposed illumination loss allows explicit control over the lighting in a view-consistent manner.

2 Related Work

2D Generation and Editing Research on unconditional or text-driven image generation has made fruitful progress. Generative Adversarial Networks (GANs) [14] contribute to the first revolutionary 2D image generative methods, among which StyleGAN [22] and its variants [20, 23] stand out due to

the expressive and well-disentangled latent spaces. StyleGAN-based image editing requires either an encoder trained to map a given image to the latent space [2, 35], or specifying latent update direction which requires explicit ground-truth annotations [1, 16, 52, 57]. Diffusion models [49, 50] represent another class of generative models that enables text-driven, photorealistic and highly diverse image generation. State-of-the-art text-to-image synthesis methods contribute effective mechanisms to guide samples toward semantics: classifier-free guidance [17, 33] generates images with or without class information during model training; CLIP guidance [10, 38, 61] where CLIP [41] trained on 400 million image-text pairs spearheaded cross-modal representation learning in modern vision-language tasks. Leveraging the rich joint embedding spaces of CLIP and expressiveness of diffusion models, stable diffusion [43] is arguably the best text-to-image model to date, synthesizing images of vast domains and styles based on a text prompt. Thus we adopt stable diffusion as the guidance model in our 3D approach.

3D Generation and Editing NeRF [32], an implicit neural representation, has become the dominating modern approach for 3D generation due to its continuity, differentiability, compactness, and quality of novel-view synthesis over mesh and point cloud. GRAF [45] combines implicit neural rendering with GAN for generalizable NeRF. PiGAN [5] utilizes SiREN [48] to condition the implicit neural radiance field on the latent space. Although guaranteed with 3D consistency, volumetric rendering requires heavy computation. With limited computation, the image quality of these methods is still not comparable to those produced by current state-of-the-art 2D GANs. Thus, many recent approaches adopt hybrid structures. StyleNeRF [15] applies volume rendering in the early feature maps in low resolution, followed by upsampling blocks to generate high-resolution images. However, a regularizer based on NeRF is required to ensure 3D consistency during upsampling. Instead of using volume rendering in early layers, EG3D [6] performs the operation on a relatively high-resolution feature map using a hybrid representation for 3D features generated by StyleGAN2 [23] backbone, named tri-plane, which is capable of incorporating more information than an explicit structure such as voxel. StyleSDF [36] shares a similar spirit but uses SiREN [48] for its mapping network, with the mapped result used as the input feature map followed by a style-based generator for upsampling.

Attempts have been made to generate 3D objects using diffusion models. Rodin [55], Realfusion [31] and other diffusion-based 3D reconstruction methods have demonstrated the feasibility of constructing a face or other object from a single view image. However, the results are not adequately realistic with limited generalizability due to the scarcity of 3D data.

Illumination Control 3D editable lighting on NeRF is a highly desirable feature. Image relighting methods can be roughly categorized into two groups: one involves estimating 3D face information such as 3DMM coefficients [4], albedo, and surface normals, and combining this information with a target lighting condition represented by spherical harmonics (SH) coefficients to generate relit images, such as [18, 30, 12]. Although guaranteed with 3D consistency, these methods have limited editability as they cannot easily accommodate changes in the original model, such as wearing glasses or different hairstyles. The other approach involves leveraging 2D/3D generative models to control illumination in the latent space, such as [3, 26]. While this approach is conducive to some editability, illumination can only be implicitly controlled by latent codes, that is, environmental lighting cannot be directly controlled by e.g., SH coefficients or cube mapping. Our method on the other hand is amenable to both light editing means, bypassing any intrinsic separation which is not necessary in FDNeRF.

3 Method

3.1 EG3D and $\mathcal{W}^+$ Space

Our method utilizes trained EG3D generator [6] with its respective latent space. From initial latent code $z \in \mathcal{R}^{512}$, a mapping network $\mathcal{M}$ maps z to w in the space named $\mathcal{W}$, where $w \in \mathcal{R}^{1\times512}$. During training and generation, the w code will be utilized to modulate all convolution layers as in StyleGAN2 [23]. The generated features will be reshaped into three orthogonal feature plans (F_{xy}, F_{yz}, F_{xz}), where each of them has the resolution of $N \times N \times C$. Given a camera pose, an augmented feature can be rendered using volume rendering as in [32], which will then be upsampled by the super-resolution blocks. The rendering process acts as an inductive bias that enforces view-consistent results.

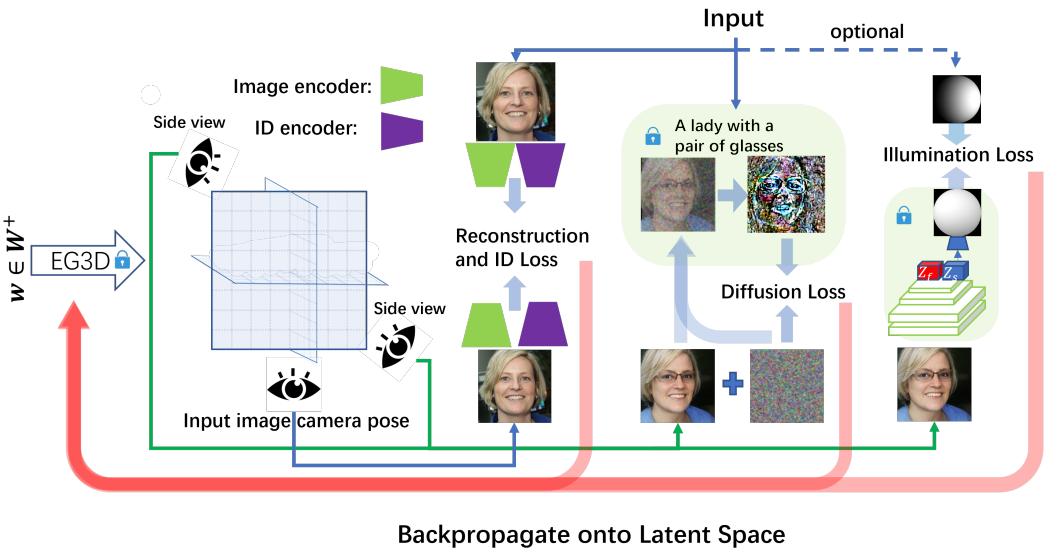

Figure 2: **FDNeRF structure**. The Latent 512 scalars are initialized as the mean of sampled $\mathcal{W}^+$ codes. The model computes the Reconstruction Loss and Identity Loss from the input image's camera view. From other side views, given a text prompt, the model computes the Diffusion Loss by assessing the discrepancy between the predicted noise and random noise. The Illumination Loss is then computed by comparing the estimated SH coefficients of the rendered side-view image to the target SH coefficients. These Losses are then utilized to update the Latent 512 scalars iteratively through a carefully designed differentiable model via back-propagation.

The inversion process inverses an input image to the latent space, such that the inverted code w' can faithfully reconstruct the given input. As shown in 2D GAN inversion [42], the quality of reconstruction is better when the inversion is conducted on $\mathcal{W}^+$ space, where its latent codes $w^+ \in \mathcal{R}^{L\times 512}$ are used to separately modulate all L convolution blocks. Thus, the reconstruction and editing process of our method operates on the $\mathcal{W}^+$ space.

3.2 Formulation

Figure 2 summarizes our method which takes as input a single image x, text prompt y, and target lighting denoted by SH coefficients l and outputs a reconstructed 3D face NeRF. Inspired by high-fidelity face NeRF reconstruction such as EG3D inversion [6], we guide the face NeRF reconstruction by the reconstruction Loss L_R, Identity Loss $\mathcal{L}_{ID}$ (Section 3.3). To perform editing, a Diffusion Loss $\mathcal{L}_D$ (section 3.4) based on the trained text-conditioned latent diffusion model is incorporated. Our method allows explicit illumination control thanks to our Illumination Loss $\mathcal{L}_{IL}$ (section 3.5). Given a latent code randomly sampled near the mean value of the latent space $w \in \mathcal{W}+$, we solve the following optimization problem:

$$\arg\min_{w\in\mathcal{W}+} \lambda_{ID}\mathcal{L}_{ID}(G(w,c),x)+\lambda_R\mathcal{L}_R(G(w,c),x)+\lambda_D\mathcal{L}_D(G(w,c_s),y)+\lambda_{IL}\mathcal{L}_{IL}(G(w,c_s),l_{c_s}). \tag{1}$$

where the $G(\cdot,\cdot)$ is the EG3D generator, c is the camera pose of the input image (which can be estimated by [13]), c_s is the random side camera pose. l_{c_s} is the target illumination which varies with c_s. λ_{ID}, λ_R, λ_D and λ_{IL} are weights of the corresponding loss functions. We optimize the w iteratively through gradient descent by back-propagating the gradient of the objective eq. (1) through these four weighted differentiable loss functions.

3.3 Reconstruction Loss and Identity Loss

The desired editing should preserve the background and identity of the input image, and hence we adopt the Reconstruction Loss and Identity Loss. Given input image x and rendered image $G(w,c)$, we utilize VGG16 image encoder $V(\cdot)$ [47] to extract the image features and construct the

Reconstruction Loss as following:

$$\mathcal{L}_R(G(w,c),x) = ||V(G(w,c)) - V(x)||_2^2 \tag{2}$$

For human faces, we adopt the same Identity Loss as [42]:

$$\mathcal{L}_{ID}(G(w,c),x) = 1 - \langle R(x), R(G(w,c))\rangle \tag{3}$$

where R is the pretrained ArcFace [11] network. Note that VGG16 image encoder and ArcFace $\mathcal{L}_R$ and $\mathcal{L}_{ID}$ are only used to measure the difference between the input image and the rendered image from input camera's view to avoid the misalignment due to mismatched viewing directions. See our ablation study for an insightful analysis of the interaction between the Reconstruction and Identity losses.

3.4 Diffusion Loss

To utilize the trained 2D diffusion model, we modify the Score Distillation Sampling (SDS) from DreamFusion [40] as our Diffusion Loss function. The loss connects the x' rendered from a sampled camera pose c_s with the denoising prediction conditioned on the editing prompt y. The denoising process can be written as:

$$\mathcal{L}_D(x' = G(w,c_s), y) = \mathbb{E}_{\varepsilon(x'),y,t,\epsilon}\left[\| \hat{\epsilon_\phi}(\mathbf{z_t}; y, t) - \epsilon \|_2^2\right] \tag{4}$$

where $\varepsilon(\cdot)$ is the image encoder that encodes our rendered image $x' = G(w, c_s)$ to the latent space of the diffusion model, denoted as z. Here z_t is the noisy version of z at time-step t, $\hat{\epsilon_\phi}$ is the frozen denoising network of the trained diffusion model, where we sample $t \sim \mathcal{U}(0.02, 0.98)$ to avoid very high and low noise levels; $\epsilon \sim \mathcal{N}(0, I)$, and its effect on input latent varies with time-step t, which is same as done in Dreamfusion [40].

Notably, the camera pose c_s will be resampled randomly in each optimization iteration to ensure 3D view consistency. Unlike DreanFusion, our text prompt y is view independent, since we have the identity and reconstruction losses to constrain the optimization process, and that the EG3D generator has underlying 3D information in contrast to randomly initialized NeRF. Therefore, our diffusion loss back-propagates to the latent scalars, where the gradient of $\mathcal{L}_D$ is given by

$$\nabla_w \mathcal{L}_D(x' = G(w,c_s), y) \triangleq \mathbb{E}_{\varepsilon,t,\varepsilon}\left[w(t)(\hat{\varepsilon_t}(\mathbf{z_t}; y, t) - \epsilon)\frac{\partial x'}{\partial w}\right] \tag{5}$$

As in Dreamfusion [40], $w(t)$ absorbs the coefficient of the forward process, and we ignore the term $\frac{\partial \hat{\varepsilon_t}(\mathbf{z_t}; y, t)}{\partial \mathbf{z_t}}$ as the diffusion model is frozen.

3.5 Illumination Loss

We utilize [64] to construct our Illumination Loss, where the hourglass network can be denoted as:

$$\mathbf{L}_s^*, \mathbf{I}_t^* = \mathbf{Hn}(\mathbf{L}_t, \mathbf{I}_s) \tag{6}$$

where $\mathbf{L}_t$ and $\mathbf{L}_s^*$ are respectively the target SH lighting and the estimated SH lighting of the input image $\mathbf{I}_s$, and $\mathbf{I}_t^*$ is the relit image under the target SH lighting. Here, we only use $\mathbf{L}_s^*$ to compute the L_1 loss with $\mathbf{L}_t$, and $\mathbf{I}_t^*$ will be ignored, i.e., $\mathbf{L}_s^* = \mathbf{Hn}'(\mathbf{I}_s)$. Thus, our Illumination Loss is:

$$\mathcal{L}_{IL}(G(w,c_s), l_{c_s}) = \| \mathbf{Hn}'(G(w,c_s)) - l_{c_s} \|_1 \tag{7}$$

During the optimization process of w^*, differentiability is crucial. Given $G(w, c_s)$ is differentiable with respect to w, the differentiability of $\mathbf{Hn}'(G(w, cs))$ with respect to $G(w, c_s)$ guarantees the back-propagation onto the w. Therefore, we replace the PyTorch in-place operations and other numpy operations in the $\mathbf{Hn}'$ model with equivalent differentiable PyTorch tensor operations. As verified by experiments, our modified $\mathbf{Hn}'$ achieves the same performance as the original model while being multi-view consistent as we render the output.

4 Experiments

In this section, we present our editing results and compare them with representative methods, emphasizing FDNeRF's disentanglement capability which is conducive to editing control, its flexibility in utilizing a text-guided diffusion model, as well as the multi-view consistency in illumination control. Furthermore, our method can be migrated to other data domains with trained generative models.

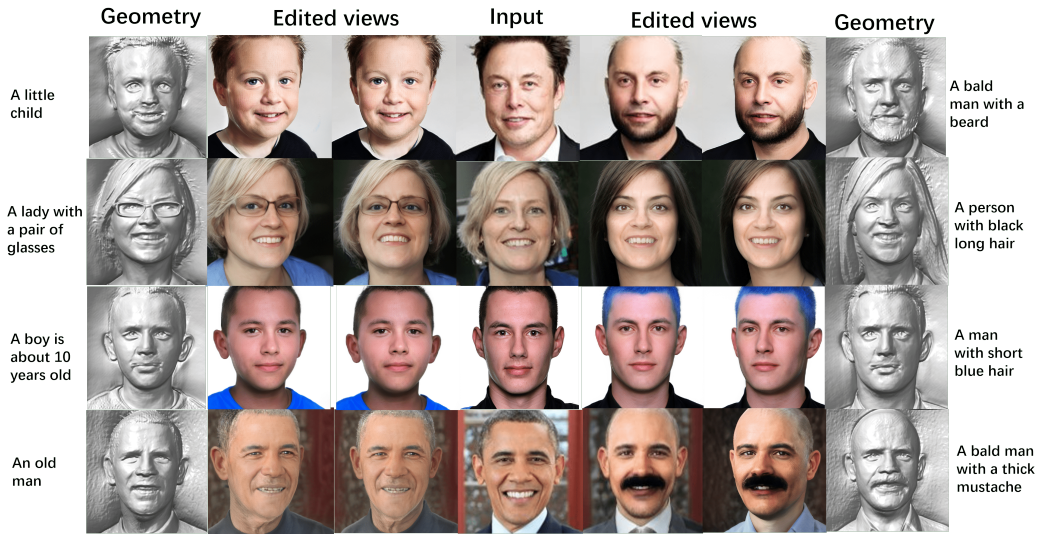

Figure 3: **More FDNeRF results.** The middle column shows the input images, while the left and right halves of the images in the other two columns show the results of text prompts editing on the left and right, respectively. Additionally, the first and last columns showcase the corresponding geometries.

4.1 3D Editing from Single Image

Although diffusion models can generate diverse and realistic images in 2D, the lack of large and high-quality 3D datasets limits the performance of current diffusion models. On the other hand, the adversarial learning mechanism together with inductive bias of 3D rendering enables GAN to produce high-quality 3D results. From GAN's smooth latent space, our method finds the suitable latent code whose semantic information matches the editing demand. Figure 3 demonstrates the detailed editing results of FDNeRF.

4.2 Comparison

Comparison with representative editing methods For effective and easy editing, the underlying latent space should be smooth and semantically meaningful. Compared with the latent space of 2D GANs, disentangling in 3D is much harder. InterfaceGAN [46] seeks a hyper-plane in the latent space with pre-trained classifiers, which makes editing feasible by interpolating latent codes in the direction orthogonal to the hyper-plane. However, the assumption that a good hyper-plane exists which is well-behaved for linear interpolation is only valid when the latent space is highly disentangled. As shown in Figure 4, the underlying latent space in 3D is not well disentangled. Thus the interpolation (induced by editing) can interfere and affect other irrelevant attributes not to be edited. We also compare with the semantic-based method FENeRF [51], whose editing is performed by manually editing the semantic map. However, this representation limits editable attributes, especially on semantically complex attributes such as age and gender. In addition, we compare with language-guided StyleClip [38] using the same latent space. Our method achieves better editing quality conditioned on the same text prompt, indicating the superiority of utilizing the diffusion model as guidance.

Illumination Comparison There have been attempts in 2D to control illumination, either implicitly or explicitly. To produce results compatible in 3D, a straightforward approach is rendering multi-view images from a given NeRF, followed by using the projected lighting direction for each view to render the edited NeRF. However, as shown in Figure 5, without the constraint of 3D rendering, inconsistent illumination is easily observed, indicating imperfect alignment using 2D methods. With explicit 3D control, FDNeRF directly manipulates latent code in the disentangled space, capable of rendering consistent and realistic lighting results.

Migration to Other Data Domain FDNeRF is not limited to human faces: the semantically diverse and rich text-guidance diffusion model can be used to directly edit other data domains. As

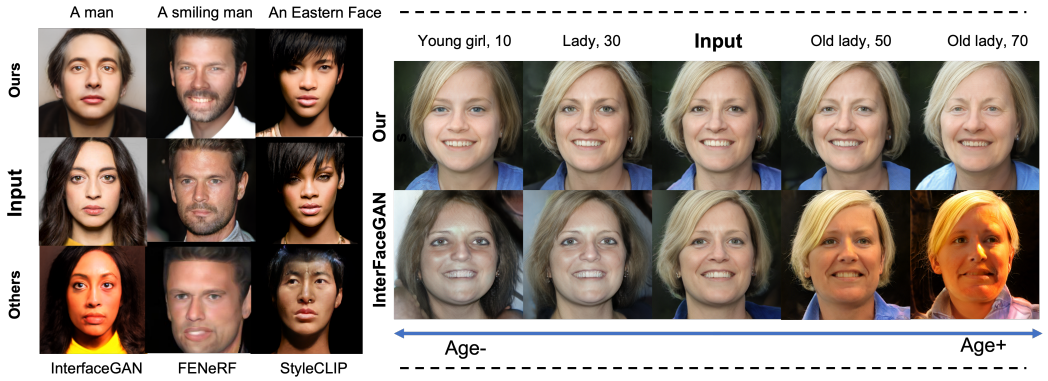

Figure 4: **Editing Comparison with representative methods**. We compare with classifier-based InterfaceGAN [46], semantic edited FENeRF [51] and language guided StyleClip [38] with EG3D. Images on the left side are editing results of gender, smile, and "An Eastern face" respectively. The right images are age editing comparison with [46], and our input text prompt are "A *XX* is about *YY* years old", where *XX* and *YY* are depicted above. The comparison illustrates the flexibility and high-fidelity editing capability of FDNeRF.

Figure 5: **Illumination Comparison**. The images in the first row are relit images generated by DPR [64]. The images in the second row are relit images generated by our method. Upon comparing the four red boxes, our method is observed to produce fewer artifacts and more realistic results. Furthermore, the images in the third row demonstrate that our method can not only edit face NeRF using text prompts but also allows explicit control of lighting.

shown in Figure 1, we perform editing on GANs trained on Cats and Cars, noting the reconstruction loss is also universal to all data domains.

4.3 Text-conditioned 3D Generation

In addition to editing, FDNeRF can generate high-quality 3D models given a text prompt. Similar to Dreamfusion [40], we guide the generation process under the iterative supervision of a trained diffusion model. However, their unconstrained optimization with random NeRF initialization lacks the ability to generate realistic results. fig. 6 shows some generated high-quality examples by our methods, including examples from other data domains.

Latent Regularization Although the latent space is smooth, generation from rare-sampled latent codes may produce unrealistic results. Latent samples around mean latent code tends to give better outputs (truncation trick in StyelGAN2 [23]). In text-conditioned 3D generation, the optimized latent code is more likely to deviate from the latent mean since there is no reconstruction or identity restriction. Therefore, we add a latent regularization to encourage results around the mean, formulated

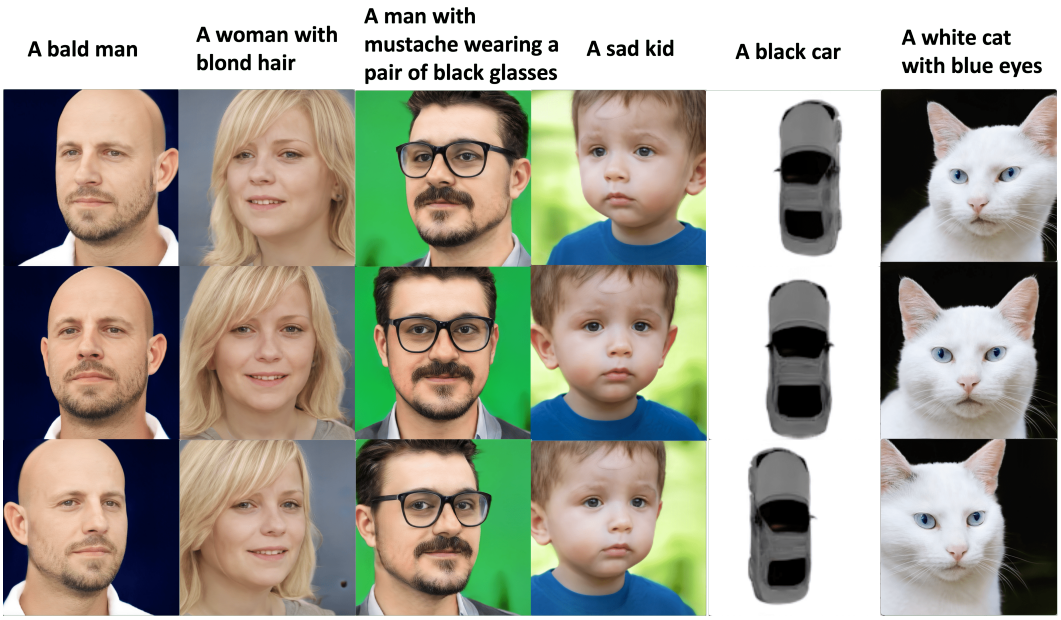

Figure 6: **Text-driven 3D Generation**. These images are generated and conditioned solely on text prompts. We utilize the trained EG3D generators [6] from different data domains.

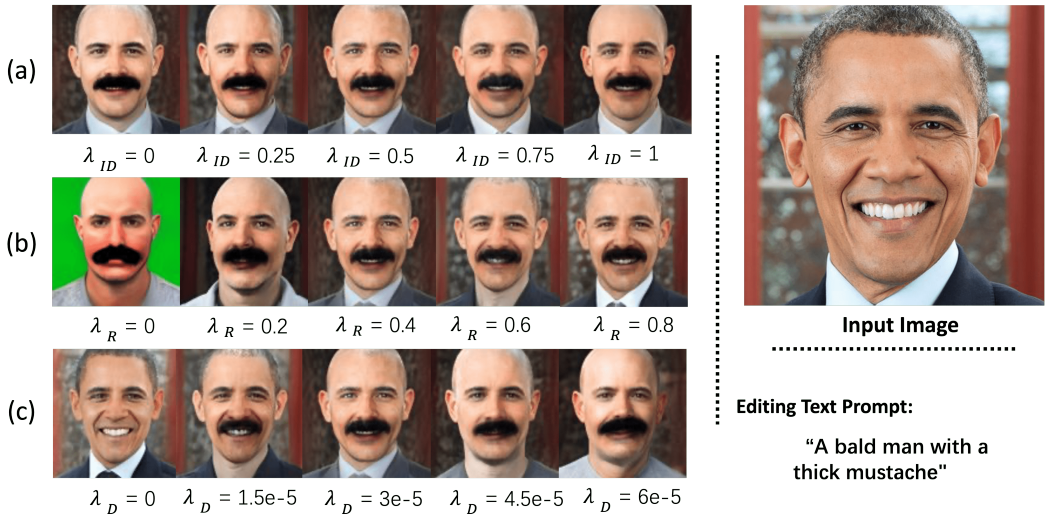

Figure 7: **Ablation studies on λ_{ID}, λ_R, and λ_D**. The input text prompt is "A bald man with a thick mustache". The central image in each row is the same, with the setting (λ_{ID}, λ_R, λ_D) = (0.5, 0.4, 3×10^{-5}). Each row shows the impact of changing one of the weights associated with a particular loss term.

as:

$$\mathcal{L}_{regu}(w, \bar{w}) = \lambda_{regu}\|w - \bar{w}\|_2^2 \tag{8}$$

where $\bar{w}$ is the sample mean of latent space. This regularization is equivalent to constraining the feasible space around the mean latent code (Lagrange Multiplier).

4.4 Ablation Study

To investigate the influence of different losses on the generated NeRF, we conduct ablation studies on:

(a) Weight of Identity Loss λ_{ID};

(b) Weight of Reconstruction Loss λ_R;

(c) Weight of Diffusion Loss λ_D.

fig. 7 shows the frontal-view rendering of the resulting NeRF in our ablation studies. The input text prompt is "A bald man with a thick mustache". The central image in each row is the same, adopting (λ_{ID}, λ_R, λ_D) = (0.5, 0.4, 3×10^{-5}). In each row, only one weight is changed to show the effect of the corresponding loss term.

By observing the rendering results in row (a) and row (b) in fig. 7, we notice that results with either larger λ_{ID} or larger λ_R look more similar to the input person, indicating that both Identity Loss and Reconstruction Loss help preserve the person's identity. However, the two losses operate in different ways.

Identity Loss is calculated using pre-trained ArcFace [11] network as described in section 3.3. ArcFace [11] trains DCNNs for face recognition which maps the face image into a feature with small intra-class but large inter-class distance, which means a person with different expressions, hairstyles, mustache, or subtle wrinkles, should have a similar identity score, thus low Identity Loss in our case. Therefore, Identity Loss would not encode details such as hairstyles and subtle wrinkles, but help to preserve large-scale identity-specific information, such as the shape of the head, nose, mouth, eyes, eyebrows, and any apparent wrinkles.

Instead, Reconstruction Loss calculated by eq. (2) takes all pixels into consideration, thus trying to preserve all kinds of features indiscriminately. A large value of λ_R results in a face NeRF similar to the input image. However, this dilutes the guidance effect from the text prompt's attempt to edit the NeRF by removing hair and adding a mustache in fig. 7. From the right two images of row (b) in fig. 7, we observe that the identity still has apparent hair, since the large weight of Reconstruction Loss forces the face NeRF to be as similar as possible to the input image.

This is a trade-off between fidelity to details in the input image versus fidelity to the text prompt. As a result, our joint utilization of Identity Loss and Reconstruction Loss can significantly alleviate the problem, where Identity Loss can preserve the identity information in a less contradictory way in the presence of text guidance.

Results in row (c) of fig. 7 shows the effect of λ_D. A larger value of λ_D enhances guidance from the diffusion model, while a smaller value of λ_R produces higher-fidelity results to the input image. Users could choose suitable value of λ_{ID}, λ_R, and λ_D based on the specific task. More ablation studies can be found in the supplementary materials.

5 Conclusion and Discussion

We propose FDNeRF, a new generative method to reconstruct high-quality Face NeRFs from a single image, with semantic prompt editing and relighting capitalizing on recent stable diffusion contributions. Extensive experiments validate our significant improvement over state-of-the-art 3D face reconstruction and editing methods. The proposed FDNeRF is readily applicable to many real-world applications such as 3D face manipulation, which, however, might be used unethically. Also, the upper bound of our performance is limited by the chosen GANs or diffusion models. We leverage EG3D [6], which is trained on real-world datasets, thus our generation results are realistic but confined to real-world faces or objects, FDNeRF can go beyond faces and can be extended to a generic NeRF generation and editing template, where different 3D generators can replace EG3D [6] to produce NeRFs in various domains with ease.

References

[1] Abdal, R., Zhu, P., Mitra, N.J., Wonka, P., 2020. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. CoRR abs/2008.02401. URL: https://arxiv.org/abs/2008.02401, arXiv:2008.02401.

[2] Alaluf, Y., Patashnik, O., Cohen-Or, D., 2021. Only a matter of style: Age transformation using a style-based regression model. CoRR abs/2102.02754. URL: https://arxiv.org/abs/2102.02754, arXiv:2102.02754.

[3] Bhattad, A., Forsyth, D., 2023. Stylitgan: Prompting stylegan to generate new illumination conditions, in: arXiv.

[4] Blanz, V., Vetter, T., 1999. A morphable model for the synthesis of 3d faces, in: Proceedings of the 26th annual conference on Computer graphics and interactive techniques, pp. 187–194.

[5] Chan, E., Monteiro, M., Kellnhofer, P., Wu, J., Wetzstein, G., 2020. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis, in: arXiv.

[6] Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., Mello, S.D., Gallo, O., Guibas, L.J., Tremblay, J., Khamis, S., Karras, T., Wetzstein, G., 2021. Efficient geometry-aware 3d generative adversarial networks. CoRR abs/2112.07945. URL: https://arxiv.org/abs/2112.07945, arXiv:2112.07945.

[7] 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.

[8] Chen, H., Tretschk, E., Stuyck, T., Kadlecek, P., Kavan, L., Vouga, E., Lassner, C., 2022. Virtual elastic objects. CoRR abs/2201.04623. URL: https://arxiv.org/abs/2201.04623, arXiv:2201.04623.

[9] 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.

[10] Crowson, K., Biderman, S., Kornis, D., Stander, D., Hallahan, E., Castricato, L., Raff, E., 2022. Vqgan-clip: Open domain image generation and editing with natural language guidance. arXiv:2204.08583.

[11] Deng, J., Guo, J., Xue, N., Zafeiriou, S., 2019a. Arcface: Additive angular margin loss for deep face recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[12] Deng, Y., Yang, J., Chen, D., Wen, F., Tong, X., 2020. Disentangled and controllable face image generation via 3d imitative-contrastive learning, in: IEEE Computer Vision and Pattern Recognition.

[13] Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., Tong, X., 2019b. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set, in: IEEE Computer Vision and Pattern Recognition Workshops.

[14] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial networks. arXiv:1406.2661.

[15] Gu, J., Liu, L., Wang, P., Theobalt, C., 2021. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. CoRR abs/2110.08985. URL: https://arxiv.org/abs/2110.08985, arXiv:2110.08985.

[16] Härkönen, E., Hertzmann, A., Lehtinen, J., Paris, S., 2020. Ganspace: Discovering interpretable GAN controls. CoRR abs/2004.02546. URL: https://arxiv.org/abs/2004.02546, arXiv:2004.02546.

[17] Ho, J., Salimans, T., 2022. Classifier-free diffusion guidance. arXiv:2207.12598.

[18] Hou, A., Sarkis, M., Bi, N., Tong, Y., Liu, X., 2022. Face relighting with geometrically consistent shadows, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[19] Jiang, Y., Jiang, S., Sun, G., Su, Z., Guo, K., Wu, M., Yu, J., Xu, L., 2022. Neuralhofusion: Neural volumetric rendering under human-object interactions. arXiv:2202.12825.

[20] Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T., 2021. Alias-free generative adversarial networks. CoRR abs/2106.12423. URL: https://arxiv.org/abs/2106.12423, arXiv:2106.12423.

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

[22] Karras, T., Laine, S., Aila, T., 2018b. A style-based generator architecture for generative adversarial networks. CoRR abs/1812.04948. URL: http://arxiv.org/abs/1812.04948, arXiv:1812.04948.

[23] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T., 2019. Analyzing and improving the image quality of stylegan. CoRR abs/1912.04958. URL: http://arxiv.org/abs/1912.04958, arXiv:1912.04958.

[24] Kim, G., Ye, J.C., 2021. Diffusionclip: Text-guided image manipulation using diffusion models. CoRR abs/2110.02711. URL: https://arxiv.org/abs/2110.02711, arXiv:2110.02711.

[25] Kundu, A., Genova, K., Yin, X., Fathi, A., Pantofaru, C., Guibas, L., Tagliasacchi, A., Dellaert, F., Funkhouser, T., 2022. Panoptic Neural Fields: A Semantic Object-Aware Neural Scene Representation.

[26] Kwak, J.g., Li, Y., Yoon, D., Kim, D., Han, D., Ko, H., 2022. Injecting 3d perception of controllable nerf-gan into stylegan for editable portrait image synthesis, in: European Conference on Computer Vision, Springer. pp. 236–253.

[27] Li, H., Yang, X., Zhai, H., Liu, Y., Bao, H., Zhang, G., 2023. Vox-surf: Voxel-based implicit surface representation. arXiv:2208.10925.

[28] Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y., 2023. Magic3d: High-resolution text-to-3d content creation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[29] Liu, S., Zhang, X., Zhang, Z., Zhang, R., Zhu, J., Russell, B., 2021. Editing conditional radiance fields. CoRR abs/2105.06466. URL: https://arxiv.org/abs/2105.06466, arXiv:2105.06466.

[30] Liu, Y., Shu, Z., Li, Y., Lin, Z., Zhang, R., Kung, S.Y., 2022. 3d-fm gan: Towards 3d-controllable face manipulation, in: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (Eds.), Computer Vision – ECCV 2022, Springer Nature Switzerland, Cham. pp. 107–125.

[31] Melas-Kyriazi, L., Rupprecht, C., Laina, I., Vedaldi, A., 2023. Realfusion: 360 reconstruction of any object from a single image, in: CVPR. URL: https://arxiv.org/abs/2302.10663.

[32] Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R., 2020. Nerf: Representing scenes as neural radiance fields for view synthesis, in: ECCV.

[33] Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M., 2021. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. CoRR abs/2112.10741. URL: https://arxiv.org/abs/2112.10741, arXiv:2112.10741.

[34] Niemeyer, M., Geiger, A., 2020. Giraffe: Representing scenes as compositional generative neural feature fields.

[35] Nitzan, Y., Bermano, A., Li, Y., Cohen-Or, D., 2020. Disentangling in latent space by harnessing a pretrained generator. CoRR abs/2005.07728. URL: https://arxiv.org/abs/2005.07728, arXiv:2005.07728.

[36] Or-El, R., Luo, X., Shan, M., Shechtman, E., Park, J.J., Kemelmacher-Shlizerman, I., 2021. Stylesdf: High-resolution 3d-consistent image and geometry generation. CoRR abs/2112.11427. URL: https://arxiv.org/abs/2112.11427, arXiv:2112.11427.

[37] Ost, J., Mannan, F., Thuerey, N., Knodt, J., Heide, F., 2020. Neural scene graphs for dynamic scenes.

[38] Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D., 2021. Styleclip: Text-driven manipulation of stylegan imagery. arXiv:2103.17249.

[39] Peng, S., Dong, J., Wang, Q., Zhang, S., Shuai, Q., Zhou, X., Bao, H., 2021. Animatable neural radiance fields for modeling dynamic human bodies. arXiv:2105.02872.

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

[41] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning transferable visual models from natural language supervision. CoRR abs/2103.00020. URL: https://arxiv.org/abs/2103.00020, arXiv:2103.00020.

[42] Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., Cohen-Or, D., 2021. Encoding in style: a stylegan encoder for image-to-image translation, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[43] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2021. High-resolution image synthesis with latent diffusion models. arXiv:2112.10752.

[44] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M., 2022. Photorealistic text-to-image diffusion models with deep language understanding. arXiv:2205.11487.

[45] Schwarz, K., Liao, Y., Niemeyer, M., Geiger, A., 2020. GRAF: generative radiance fields for 3d-aware image synthesis. CoRR abs/2007.02442. URL: https://arxiv.org/abs/2007.02442, arXiv:2007.02442.

[46] Shen, Y., Yang, C., Tang, X., Zhou, B., 2020. Interfacegan: Interpreting the disentangled face representation learned by gans. arXiv:2005.09635.

[47] Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.

[48] Sitzmann, V., Martel, J.N.P., Bergman, A.W., Lindell, D.B., Wetzstein, G., 2020. Implicit neural representations with periodic activation functions. CoRR abs/2006.09661. URL: https://arxiv.org/abs/2006.09661, arXiv:2006.09661.

[49] Sohl-Dickstein, J., Weiss, E.A., Maheswaranathan, N., Ganguli, S., 2015. Deep unsupervised learning using nonequilibrium thermodynamics. CoRR abs/1503.03585. URL: http://arxiv.org/abs/1503.03585, arXiv:1503.03585.

[50] Song, Y., Ermon, S., 2019. Generative modeling by estimating gradients of the data distribution. CoRR abs/1907.05600. URL: http://arxiv.org/abs/1907.05600, arXiv:1907.05600.

[51] Sun, J., Wang, X., Zhang, Y., Li, X., Zhang, Q., Liu, Y., Wang, J., 2021. Fenerf: Face editing in neural radiance fields. CoRR abs/2111.15490. URL: https://arxiv.org/abs/2111.15490, arXiv:2111.15490.

[52] Wang, B., Ponce, C.R., 2021. The geometry of deep generative image models and its applications. CoRR abs/2101.06006. URL: https://arxiv.org/abs/2101.06006, arXiv:2101.06006.

[53] Wang, C., Chai, M., He, M., Chen, D., Liao, J., 2021a. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. CoRR abs/2112.05139. URL: https://arxiv.org/abs/2112.05139, arXiv:2112.05139.

[54] Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W., 2021b. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. CoRR abs/2106.10689. URL: https://arxiv.org/abs/2106.10689, arXiv:2106.10689.

[55] Wang, T., Zhang, B., Zhang, T., Gu, S., Bao, J., Baltrusaitis, T., Shen, J., Chen, D., Wen, F., Chen, Q., Guo, B., 2022. Rodin: A generative model for sculpting 3d digital avatars using diffusion. arXiv:2212.06135.

[56] Weng, C.Y., Curless, B., Srinivasan, P.P., Barron, J.T., Kemelmacher-Shlizerman, I., 2022. Humannerf: Free-viewpoint rendering of moving people from monocular video. arXiv:2201.04127.

[57] Xu, Y., Yin, Y., Jiang, L., Wu, Q., Zheng, C., Loy, C.C., Dai, B., Wu, W., 2022. TransEditor: Transformer-based dual-space GAN for highly controllable facial editing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

[58] Yang, B., Zhang, Y., Xu, Y., Li, Y., Zhou, H., Bao, H., Zhang, G., Cui, Z., 2021. Learning object-compositional neural radiance field for editable scene rendering.

[59] Yang, G., Vo, M., Neverova, N., Ramanan, D., Vedaldi, A., Joo, H., 2023. Banmo: Building animatable 3d neural models from many casual videos. arXiv:2112.12761.

[60] Yariv, L., Gu, J., Kasten, Y., Lipman, Y., 2021. Volume rendering of neural implicit surfaces. CoRR abs/2106.12052. URL: https://arxiv.org/abs/2106.12052, arXiv:2106.12052.

[61] Yu, Y., Zhan, F., Wu, R., Zhang, J., Lu, S., Cui, M., Xie, X., Hua, X.S., Miao, C., 2022. Towards counterfactual image manipulation via clip. arXiv:2207.02812.

[62] Yuan, W., Lv, Z., Schmidt, T., Lovegrove, S., 2021. Star: Self-supervised tracking and reconstruction of rigid objects in motion with neural rendering, pp. 13144–13152.

[63] Zheng, Z., Huang, H., Yu, T., Zhang, H., Guo, Y., Liu, Y., 2022. Structured local radiance fields for human avatar modeling. arXiv:2203.14478.

[64] Zhou, H., Hadap, S., Sunkavalli, K., Jacobs, D.W., 2019. Deep single portrait image relighting, in: International Conference on Computer Vision (ICCV).

A 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 [21], AFHQv2 [9] and ShapeNet [7] 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.

B More Results

As a supplement to Figure 1 and Figure 3 in our main paper, fig. 8 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.

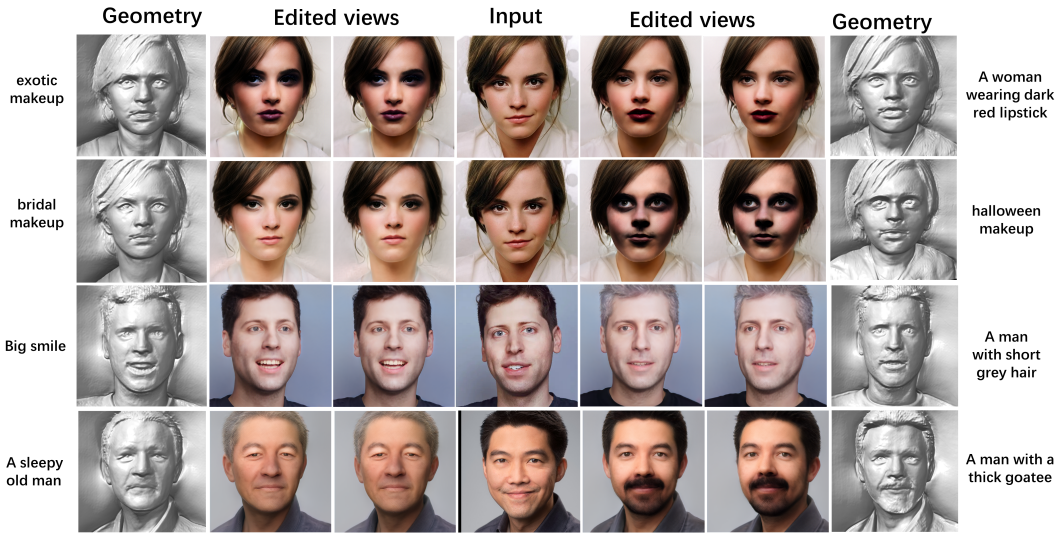

Figure 8: **More FDNeRF results.** The middle column shows the input images, while the left and right halves of the images in the other two columns show the results of text prompts editing on the left and right, respectively. Additionally, the first and last columns showcase the corresponding geometries.

Text-conditioned Generation Comparison Our attempts to compare with Dreamfusion[40] (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.

C Ablation Study on Editing Prompts

We investigate the influence of the input text prompt in this section. In fig. 8, and Figure 1 and 3 in our main paper, we mainly show NeRF editing results when the input text prompt is a short sentence

Figure 9: **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.

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. 9 shows the ablation results. The right 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. 9. 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. 10, 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. 10. "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.

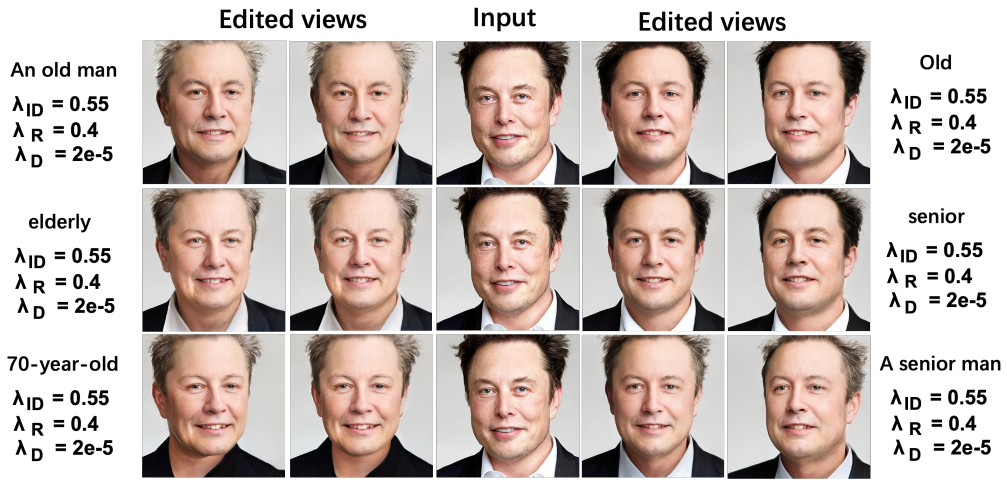

Figure 10: **Ablation study on synonyms.** Here are results generated with various Synonyms, which could affect the results. The contest is important for synonyms to eliminate ambiguity, as the case of "senior" and "A senior man".