NeRFFaceEditing: Disentangled Face Editing in Neural Radiance Fields

Kaiwen Jiang
Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences and Beijing Jiaotong University
China
kevinjiangedu@gmail.com

Shu-Yu Chen
Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences
China
chenshuyu@ict.ac.cn

Feng-Lin Liu
Beijing Key Laboratory of Mobile Computing and Pervasive Device,Institute of Computing Technology, CAS and University of Chinese Academy of Sciences
China
liufenglin21s@ict.ac.cn

Hongbo Fu
School of Creative Media, City University of Hong Kong
China
hongbofu@cityu.edu.hk

Lin Gao*
Beijing Key Laboratory of Mobile Computing and Pervasive Device, ICT, CAS and University of Chinese Academy of Sciences
China
gaolin@ict.ac.cn

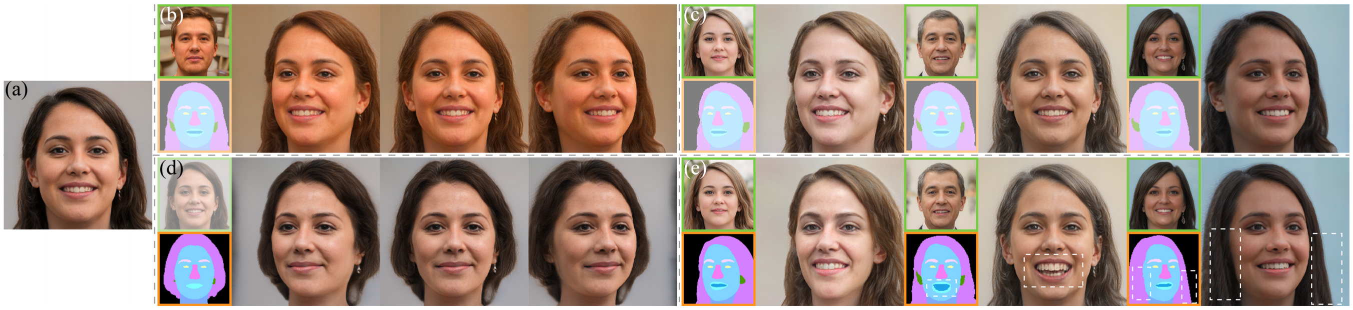

Figure 1: Our NeRFFaceEditing method allows users to intuitively edit a facial volume to manipulate its geometry and appearance guided by rendered semantic masks. Given an input sample (a), our method disentangles its geometry and appearance, and allows for one- or multi-label editing. We show a range of flexible face editing tasks that can be achieved with our unified framework: (b) changing the appearance according to a given reference sample while retaining the geometry and 3D consistency; (c) changing the appearance for different views with different reference samples while retaining the geometry; (d) editing multiple labels of the semantic mask for a certain view while keeping the appearance and 3D consistency; (e) editing both the geometry and appearance. The inputs used to control the appearance and geometry are highlighted in green and orange boxes, respectively.

ABSTRACT

Recent methods for synthesizing 3D-aware face images have achieved rapid development thanks to neural radiance fields, allowing for high quality and fast inference speed. However, existing solutions for editing facial geometry and appearance independently usually require retraining and are not optimized for the recent work of generation, thus tending to lag behind the generation process. To address these issues, we introduce NeRFFaceEditing, which enables editing and decoupling geometry and appearance in the pretrained tri-plane-based neural radiance field while retaining its high quality and fast inference speed. Our key idea for disentanglement is to use the statistics of the tri-plane to represent the high-level appearance of its corresponding facial volume. Moreover, we leverage a generated 3D-continuous semantic mask as an intermediary for geometry editing. We devise a geometry decoder (whose output is unchanged when the appearance changes) and an appearance decoder. The geometry decoder aligns the original facial volume with the semantic mask volume. We also enhance the disentanglement by explicitly regularizing rendered images with the same

*Corresponding author is Lin Gao (gaolin@ict.ac.cn).

SA '22 Conference Papers, December 6–9, 2022, Daegu, Republic of Korea
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9470-3/22/12.
https://doi.org/10.1145/3550469.3555377

appearance but different geometry to be similar in terms of color distribution for each facial component separately. Our method allows users to edit via semantic masks with decoupled control of geometry and appearance. Both qualitative and quantitative evaluations show the superior geometry and appearance control abilities of our method compared to existing and alternative solutions.

CCS CONCEPTS

• **Human-centered computing** → **Graphical user interfaces**; • **Computer systems organization** → Neural networks; • **Computing methodologies** → *Rendering*; Volumetric models.

KEYWORDS

Face editing, volume disentangling, semantic-mask-based interfaces, neural radiance fields, neural rendering

ACM Reference Format:
Kaiwen Jiang, Shu-Yu Chen, Feng-Lin Liu, Hongbo Fu, and Lin Gao. 2022. NeRFFaceEditing: Disentangled Face Editing in Neural Radiance Fields. In *SIGGRAPH Asia 2022 Conference Papers (SA '22 Conference Papers), December 6–9, 2022, Daegu, Republic of Korea.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3550469.3555377

1 INTRODUCTION

Efficiently generating consistent and high-quality 3D-aware face images is an active research topic. Many recent techniques (e.g., [Chan et al. 2021a; Deng et al. 2022; Gu et al. 2021; Or-El et al. 2021; Zhou et al. 2021]) choose to build upon Generative Adversarial [Goodfellow et al. 2014] Neural Radiance Fields (NeRF) [Mildenhall et al. 2020] and form facial volumes to generate 3D-aware high-resolution face images with 2D convolution. However, their methods lack direct control of facial geometry and some of them (e.g., [Chan et al. 2021a]) cannot control the appearance independently of the geometry, while such controls are important for applications like 3D character design, educational training, etc.

To directly control the geometry, one approach is to introduce an editing intermediary that is aligned with the facial volume. Semantic masks are suitable for 3D GANs because of their intuitiveness, ease of use, and continuity while moving the camera. FENeRF [Sun et al. 2021] has been proposed based on π-GAN to generate a volume where facial semantics and texture are spatially aligned. However, FENeRF requires time-consuming and resource-hungry retraining for enabling local editing.

On the other hand, in NeRF, to control the appearance independently of the geometry, one approach is to directly incorporate the latent code of appearance into a separated branch of color [Jang and Agapito 2021; Liu et al. 2021; Niemeyer and Geiger 2021; Schwarz et al. 2020; Sun et al. 2021]. This idea is proven to be effective and mostly applied in the coordinate-based MLP representation. Recently, many new representations have been proposed and particularly, tri-plane-based radiance fields proposed in [2021a] feature finer details, better quality, and faster inference than the coordinate-based MLP representation. However, despite the advantages of the tri-plane representation, how to decouple geometry and appearance in it while preserving its characteristics remains unexplored. Trivially extending methods working on the coordinate-based MLP representation to it might abandon its efficient tri-plane representation. Actually, there are abundant 2D style control methods (e.g., [Chen et al. 2022c; Huang and Belongie 2017]) , but their extension to 3D generation tasks remains an open problem.

In this work, we introduce NeRFFaceEditing, which relies on the pretrained tri-plane representation [Chan et al. 2021a] to address the above-mentioned limitations and aims to achieve better frontal- and side-view editing and disentanglement of geometry and appearance. In order to disentangle geometry and appearance in the tri-plane representation and inspired by adaptive instance normalization (AdaIN) [Huang and Belongie 2017], we use the mean and variance of tri-planes to represent the high-level *spatially-invariant* appearance of its corresponding *spatially-variant* facial volume. Moreover, we take the merits of the idea that colors are predicted through an additional separated branch and thus split the original decoder in EG3D into a geometry decoder to handle the geometry and an appearance decoder to handle the appearance. The geometry decoder takes in features sampled from the normalized tri-plane, while the appearance decoder takes in features sampled from the denormalized tri-plane, so that the geometry of the facial volume will not be affected when the tri-plane is stylized differently for different appearance. We choose a generated 3D-continuous semantic mask as an intermediary to enable geometry editing. The key enabler of effective editing is to make the geometry decoder directly predict both the semantic labels and densities to align the facial volume with the semantic mask volume.

To enhance disentanglement, we design a training strategy. We utilize a histogram color loss [Afifi et al. 2021] to constrain rendered images with the same appearance but different geometry to be similar in terms of color distribution for each facial component.

As shown in Fig. 1, NeRFFaceEditing allows disentangled control of geometry and appearance for better frontal- and side-view editing, enabled by the shared geometry space between the facial volume and the semantic mask volumes. Qualitative and quantitative experiments show that our approach outperforms state-of-the-art methods for various applications. To facilitate further research studies, we will release our code.

The main contributions are summarized as follows:

- We propose an AdaIN-based method and a design of decoders to decouple geometry and appearance embedded in the tri-plane and enable intuitive geometry editing by semantic masks.
- We propose a fine-tuning method to facilitate disentanglement by promoting similarities in terms of color distribution for each facial component separately in rendered images with the same appearance but different geometry.
- Our method achieves state-of-the-art 3D-aware frontal- and side-view editing based on semantic masks as well as the disentanglement of geometry and appearance proven by extensive experiments and comparisons.

2 RELATED WORK

Our work is closely related to several topics, including disentangled neural implicit representations, 3D-aware neural face image synthesis, and neural face image editing.

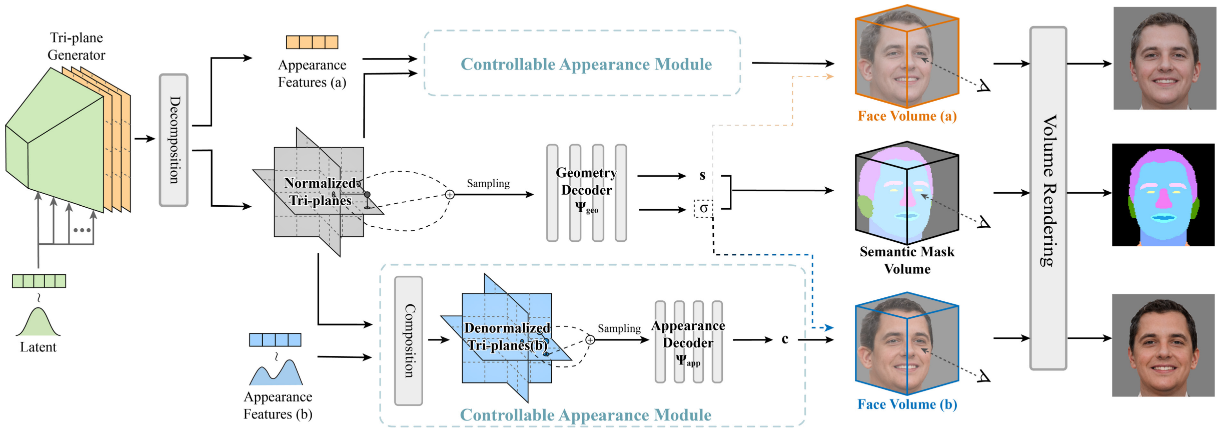

Figure 2: An overview of our framework. Our pipeline leverages the pretrained tri-plane generator to synthesize feature images. The generated feature images are then decomposed into normalized tri-planes and appearance features (a) $F^{(a)}_{\text{app}}$ through reshaping and normalization. The geometry features sampled from the normalized tri-planes are passed into a geometry decoder to output densities σ and semantic labels s, which together generate a semantic mask volume that 2D semantic masks are projected from. The normalized tri-planes, together with appearance features (a) $F^{(a)}_{\text{app}}$ and appearance features (b) $F^{(b)}_{\text{app}}$, are passed into the Controllable Appearance Module (CAM) to composite the denormalized tri-planes (a) (not shown in the figure for brevity) and (b), respectively. Features sampled from the denormalized tri-planes (a) and (b) are then passed into the same appearance decoder to output color features c in the CAM. These color features together with the same densities σ are then processed independently by a neural volume renderer to project their corresponding face volumes (a) and (b) into 2D feature images.

2.1 Disentangled Neural Implicit Representations

Neural implicit scene representation is an emerging area of study, and is originally modelled as an MLP mapping from positional-encoded coordinates to densities and colors for volume rendering. Recent developments propose to use alternative representations, including hashing tables [Müller et al. 2022], octrees [Yu et al. 2021], voxels [Sara Fridovich-Keil and Alex Yu et al. 2022], tri-planes [Chan et al. 2021a], etc. These methods feature faster inference and improved expressive power.

Given a dataset of single-view 2D facial images, to have a disentangled control of geometry and appearance, one option is to embed separated latent codes for geometry and appearance into generation (e.g., [Chan et al. 2021b; Niemeyer and Geiger 2021; Schwarz et al. 2020; Xu et al. 2021]), which, however, does not directly provide intuitive geometry manipulation. Our solution is to utilize the tri-plane representation, whose geometry and appearance are decoupled through an AdaIN-based method and the original decoder in EG3D is split into an appearance decoder and a geometry decoder. The geometry decoder models both the facial volume and the semantic mask volume, with the latter providing an editing interface. Our solution further enhances the disentanglement through a training strategy.

2.2 3D-aware Neural Face Image Synthesis

In recent years, generative models like Generative Adversarial Networks (GANs) [Goodfellow et al. 2014] combined with implicit radiance fields [Mildenhall et al. 2020] have been explored to generate 3D-consistent faces from 2D images only. To generate high-resolution images, a group of methods [Gu et al. 2021; Niemeyer and Geiger 2021; Zhou et al. 2021] first output low-resolution features and then pass them into 2D convolution. However, they suffer from the low-quality geometry representations. Thus, alternative models and representations (e.g., [Chan et al. 2021a; Deng et al. 2022; Or-El et al. 2021]) have also been explored. These models can synthesize highly realistic images with geometrically-consistent fine details.

However, all the above methods cannot control geometry in an intuitive manner. In order to address this issue, many works have employed the method of embedding explicit control into the generation process. For example, CG-NeRF [Jo et al. 2021] introduces various soft conditions, including sketches as input. FENeRF [Sun et al. 2021] involves semantic masks in the generation process as output. However, the quality of their editing results still has room for improvement and a retraining is unavoidable. By comparison, Sem2NeRF [Chen et al. 2022b] encodes single-view semantic masks into the latent space of pretrained 3D GANs. Lin et al. [2022] introduce the work [Abdal et al. 2021] on pretrained 2D GANs into 3D GANs to edit the attributes of generated results semantically. IDE-3D [Sun et al. 2022], which is concurrent with our work, achieves interactive high-quality geometry editing and disentangled appearance control. It designs an encoder to facilitate the editing and splits the original tri-planes into semantic tri-planes and texture tri-planes for disentanglement. In contrast, our method extends the pretrained tri-plane-based generative model with our unique design of decoders and operations on tri-planes for enabling intuitive editing and decoupling of geometry and appearance. Our method achieves better results with another training strategy.

2.3 Neural Face Image Editing

With the introduction of GANs, various 2D-based methods and 3D-based methods have been proposed to realize face image editing under different conditions. Here, we mainly review the approaches based on 3D generative adversarial neural radiance fields.

In 3D GANs, some methods (e.g., [Athar et al. 2021; Gafni et al. 2021; Hong et al. 2021; Zheng et al. 2022; Zhuang et al. 2021]) borrow latent codes for identity, expression, etc. from 3DMM (e.g. [Li et al. 2017a; Tran and Liu 2018]). Some other methods (e.g., [Guo et al. 2021; Kania et al. 2022; Park et al. 2020, 2021; Wang et al. 2021]) use learned embeddings to capture dynamic facial actions, which, however, are hard to interpret except for the correspondence with extra information, such as audio or using an attribute regressor. FENeRF [Sun et al. 2021] incorporates semantic masks into conditional NeRF [Chan et al. 2021b; Schwarz et al. 2020] to generate editing interfaces. Similarly, our work also utilizes semantic mask for intuitive editing with a 3D GAN as the backbone. However, our method carefully preserves its original generation power and speed without requiring retraining, and enhances the disentanglement of geometry and appearance through an additional training strategy.

3 METHODOLOGY

In this section, we formalize the structure of our proposed disentanglement framework in detail, which operates on the tri-plane representation [Chan et al. 2021a]. To decouple geometry and appearance, inspired by AdaIN [Huang and Belongie 2017], we decompose the tri-plane of each sample into the *spatially-invariant* abstract appearance features, namely the mean and variance of the tri-plane, and the normalized tri-plane where *spatially-variant* specific geometry features are sampled from (as shown in Fig. 2). Furthermore, we split the original decoder in EG3D into an appearance decoder for predicting color features and a geometry decoder for predicting densities and semantic labels to enable semantic-mask guided editing (Sec. 3.2). To further ensure disentanglement, we explicitly regularize rendered images with the same appearance but different geometry to be similar in terms of color distribution for each facial component guided by the generated semantic masks (Sec. 3.3). The editing during inference will be explained in Sec. 3.4.

3.1 Preliminaries

Since our approach is built on the tri-plane representation proposed in EG3D, it is necessary to briefly summarize EG3D's pipeline of generation here. Three planes (tri-plane for short) (p_{xy}, p_{xz}, p_{yz}) are generated by StyleGAN2 [Karras et al. 2020] f from an intermediate latent code $w \in W$. For each queried 3D position $\mathbf{x} = (x, y, z)$, its corresponding feature vector $(\mathbf{F}_{xy}, \mathbf{F}_{xz}, \mathbf{F}_{yz})$ is retrieved by projecting $\mathbf{x}$ onto each of the three planes via bilinear interpolation, and is further aggregated by summation to form final features. An additional light-weight decoder network, implemented as a small MLP Φ, interprets the aggregated 3D features $\mathbf{F}(\mathbf{x})$ as color features $\mathbf{c}(\mathbf{x})$ and densities $\sigma(\mathbf{x})$. These quantities are rendered into feature images I_F, whose first three channels are extracted as rendered images I_{RGB} in a low resolution using volume rendering [Max 1995; Mildenhall et al. 2020]. The feature images I_F are later passed into a super-resolution module, which generates high-resolution images I^+_{RGB}. The details of the super-resolution module and the discriminator are omitted for simplicity since they are not our focus.

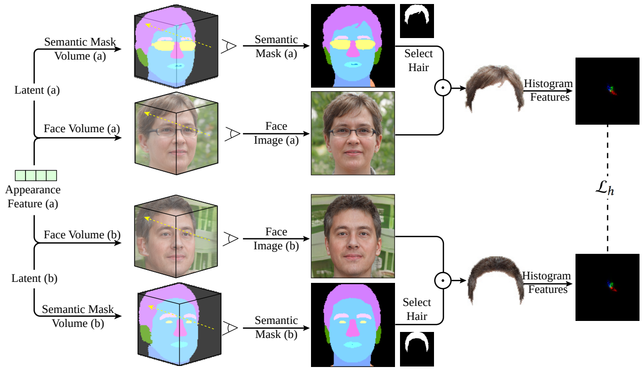

Figure 3: Illustration of promoting similarities in terms of color distribution for face images with the same appearance but different geometry. For generated face images (a) and (b) at a certain pose, their similarity is measured by the distance between histogram features for each facial component, including hair, eye, lip, skin, etc., separately, each of which is selected through the corresponding generated semantic masks. In this diagram, we take the hair as an example.

3.2 Disentanglement and Mask-guided Editing

Aiming at decoupling the geometry and appearance, one possible way is to use the cyclic swapping loss in [Chen et al. 2021]. Even though it works well in 2D frontal images, it cannot be trivially utilized in the 3D-aware facial image generation since there does not exist a consistent mapping from rendered images at any pose back to their unique appearance features.

It has been known that convolutional feature statistics can capture the style of 2D images [Gatys et al. 2016; Huang and Belongie 2017; Li and Wand 2016; Li et al. 2017b]. However, how to apply this idea to 3D-aware image generation is challenging, especially in the context of neural radiance fields where convolutional features usually do not exist. But in the case of the tri-plane representation, we are able to extend the conclusion of AdaIN [Huang and Belongie 2017] (i.e., the mean and variance can reflect the style of a 2D feature map) to tri-planes, which are practically multi-channel convolutional features. Thus, we assume that the mean and variance of the tri-plane $\mathbf{F}_{app}$ reflect its style, which represents the high-level appearance of the corresponding facial volume. The obvious benefit is that for a specific latent code w, its representation of appearance is the same for any pose P.

As in AdaIN, the style of the tri-plane is specifically controlled by the normalization and denormalization operations. We define the normalization process of the tri-plane and the denormalization process with any $\widehat{\mathbf{F}}_{app}$ respectively as:

$$\overline{p}_i = \frac{p_i - \mu_i}{\sigma_i}, \quad p'_i = \widehat{\sigma}_i \overline{p}_i + \widehat{\mu}_i, \tag{1}$$

where $i \in \{xy, xz, yz\}$, μ, σ denote the mean and the variance.

Based on the assumption that the same geometry can have various different appearances and taking the merits of the idea that colors are predicted through a separated branch, we split the original decoder Φ into a geometry decoder Ψ_{geo} and an appearance

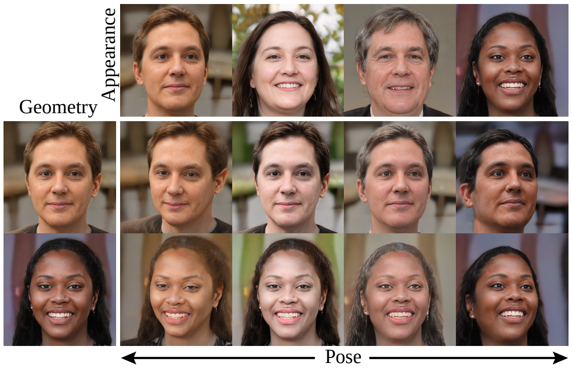

Figure 4: Style transfer. The geometry inputs (first column) are the same as the appearance inputs (first row). Each face image is generated with the the geometry reference sample in the same row and the appearance reference sample in the same column.

decoder Ψ_{app}. The former takes in the geometry features sampled from the normalized tri-plane as $\overline{\mathbf{F}}(\mathbf{x})$ and the latter takes in features sampled from the denormalized tri-plane as $\mathbf{F}'(\mathbf{x})$. We name the process of denormalizing normalized tri-planes with appearance features and decoding features sampled from denormalized tri-planes into color features as Controllable Appearance Module (CAM), as shown in Fig. 2. Thus, when the CAM is fed with the same normalized tri-planes but different appearance features, the appearance of facial volumes is changed while the geometry is unaffected.

Moreover, inspired by EditGAN [Ling et al. 2021], we enable mask-guided editing in the pretrained uneditable tri-plane representation based on the following key insight: information is concentrated in the sampled features $\mathbf{F}(\mathbf{x})$ from the tri-plane, and such information concentration is similar to the bottleneck in an encoder-decoder architecture. A lightweight decoder is responsible for transforming features in an abstract domain to various specific domains, such as densities and colors, as well as other domains like semantic labels $\mathbf{s}$. Thus, we explicitly require the geometry decoder Ψ_{geo} to predict both densities σ and semantic labels. The above pipeline can be summarized as:

$$(\sigma(\mathbf{x}), \mathbf{s}(\mathbf{x})) = \Psi_{\text{geo}}(\overline{\mathbf{F}}(\mathbf{x})), \quad (\mathbf{c}(\mathbf{x})) = \Psi_{\text{app}}(\mathbf{F}'(\mathbf{x})), \tag{2}$$

However, training another separated decoder which outputs semantic labels $\mathbf{s}$ only fails to edit precisely through GAN inversion (see the "Baseline3" of part (a) in Fig. 6). We speculate that in the case of separated branches for densities σ and semantic labels $\mathbf{s}$, the spaces between facial volumes and semantic masks are not shared or aligned and the changes in the space of semantic masks fail to propagate to the space of facial volume well. The connection between the space of facial volumes and the space of semantic masks is established through this unified decoder. Besides, empirically, we find that disentanglement is beneficial for effective editing of the hair, nose, etc. (see the "Baseline1" of part (a) of Fig. 6).

3.3 Training Process

To further improve the disentanglement, we design a specific training strategy. For simplicity, we exemplify the training with the case of low-resolution image generation. The case of high-resolution image generation only needs an additional reconstruction loss (see the supplement materials for details). At each step, we first sample a latent code w. Then, we carefully design the following losses to train our decoders Ψ_{geo} and Ψ_{app} and fine-tune the original tri-plane generator f as $\widetilde{f}$. We denote the original generation process as $G(w)$ and our new 3D-aware and disentangled generation process as $\widetilde{G}(w, \mathbf{F}_{\text{app}})$. The rendering pose is sampled from the pose distribution of face images in the dataset and omitted in all the following equations for brevity.

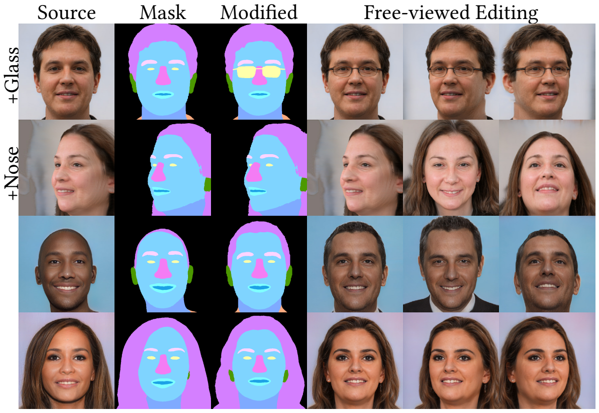

Figure 5: One- and multi-label editing. Our method enables effective and intuitive editing guided by semantic masks at frontal- and side-views. We manipulate facial attributes on the semantic map and use the optimization process described in Sec. 3.4 to obtain the modified free-view portraits.

Reconstruction Loss. To ensure high quality and diverse generation, we need to match the original generated distribution. Specifically, the training procedure is defined as:

$$\begin{aligned}(I_{RGB}, d) = G(w), \quad (I'_{RGB}, d', S') = \widetilde{G}(w, \mathbf{F}_{app}(w)),\\ \mathcal{L}_{Recon} = \lambda_1 ||I_{RGB} - I'_{RGB}|| + \lambda_2 \mathcal{L}_{VGG}(I_{RGB}, I'_{RGB})\\ + \lambda_3 E(\Theta(I_{RGB}), S') + \lambda_4 ||d - d'||,\end{aligned} \tag{3}$$

where $\mathcal{L}_{VGG}$ is the perceptual loss introduced in [Zhang et al. 2018], which measures the visual similarity between the generated images and the input images by a pretrained VGG-19 model, d and d' respectively represent the ground-truth depth image and the reconstructed depth image extracted when performing the volume rendering following [Mildenhall et al. 2020], and E denotes the pixel-wise cross-entropy loss. $\Theta(\cdot)$ stands for an off-the-shelf facial image segmentation module [Yu et al. 2018] and S' represents predicted semantic masks by the geometry decoder Ψ_{geo}. In our experiments, we empirically set $\lambda_1 = 15, \lambda_2 = 15, \lambda_3 = 1, \lambda_4 = 5$.

Part-based Histogram Loss. To enhance disentanglement explicitly, we require rendered images to be similar in terms of color distribution including the background when the CAM is fed with the same appearance features $\widehat{\mathbf{F}}_{\text{app}}$ but different normalized tri-planes. For this purpose, we utilize the histogram loss introduced in [Afifi et al. 2021], which measures the similarity among histograms representing color distributions. However, we observe that its generated results generally transfer the style but fail to capture the style in detail, especially for hairs (see the "Baseline3" of part (b) in Fig. 6). Thus, we enhance this histogram loss by performing it for each label separately guided by semantic masks (as illustrated in Fig. 3).

Specifically, we sample a batch of latent codes $w(w^{(1)}, w^{(2)}, ..., w^{(B)})$ and apply the appearance feature $\widehat{\mathbf{F}}_{\text{app}}$ of one of the latent codes $w^{(k)}$ to the normalized tri-planes generated by other latent codes for the purpose of same appearance but different geometry. Assuming that there are N classes, we define the training procedure as:

$$\begin{aligned}&(\widehat{I}_{RGB}, \widehat{d}, \widehat{S}) = \widetilde{G}(w, \widehat{\mathbf{F}}_{\text{app}})\\&\mathcal{L}_{Sim} = \lambda_5 \sum_{i=1}^{N} w_i \sum_{j=1}^{B} \mathcal{L}_h(\mathbf{H}(\hat{I}_{RGB}^{(j)} \odot M_i^{(j)}), \mathbf{H}(\hat{I}_{RGB}^{(k)} \odot M_i^{(k)})),\end{aligned} \tag{4}$$

where $\sum_{i=1}^{N} w_i = 1$, $\mathbf{H}(\cdot)$ denotes the extraction of histogram features from images, $\mathcal{L}_h$ stands for the distance among histogram features, $\hat{I}_{RGB}$ represents rendered images from $\widehat{F}_{\text{app}}$, M_i denotes the mask for label i in the semantic masks $\widehat{S}$ corresponding to $\hat{I}_{RGB}$, and B represents the batch size. In our experiments, we empirically set $\lambda_5 = 15, B = 3, k = 1$.

Our final objective $\mathcal{L}$ is simply the sum of the above two losses (as we have already weighted each term in these losses): $\mathcal{L} = \mathcal{L}_{Recon} + \mathcal{L}_{Sim}$. Minimizing $\mathcal{L}$ will lead to the optimization of three networks: $\Psi_{\text{geo}}, \Psi_{\text{app}}, \widetilde{f}$.

3.4 Editing during Inference

We can edit samples $\mathbf{I}$ at a certain pose generated by a latent code $w \in W$ space or real images. To edit given real face images $\mathbf{I}'$, we first invert the images into the $\mathcal{W}$ space as w with pivotal tuning inversion [Roich et al. 2021] following EG3D, denoted as $\mathbf{I}$. In all the editing cases, a user-edited semantic mask $\hat{S}$ is assumed to be available given an original semantic mask S.

Formally, we are seeking an editing vector $\delta w^+ \in \mathcal{W}^+$ such that $(\mathbf{I}_{\text{edited}}, \mathbf{S}_{\text{edited}}) = G'(w + \delta w^+, \mathbf{F}_{\text{app}}(w))$, in which $\mathbf{I}_{\text{edited}}$ is generated by the optimized latent code and $\mathbf{S}_{\text{edited}}$ approximates $\hat{\mathbf{S}}$. Note that the appearance feature is kept fixed during optimization to keep the appearance unchanged. As in EditGAN [Ling et al. 2021], we first define a region of interest r within which we expect the image to change due to a certain edit. A sequence of such edits can be achieved step by step and is illustrated in the supplementary.

To optimize δw^+ so that $\mathbf{S}_{\text{edited}}$ approximates $\hat{S}$ while preserving regions outside of r untouched, we use the following losses as the minimization targets:

$$\begin{aligned}\mathcal{L}_{\text{editing}}(\delta w^+) = &\mathcal{L}_{\text{VGG}}(\mathbf{I}_{\text{edited}} \odot (1-r), \mathbf{I} \odot (1-r))\\&+ \mathcal{L}_{\text{MSE}}(\mathbf{I}_{\text{edited}} \odot (1-r), \mathbf{I} \odot (1-r)) + E(\mathbf{S}_{\text{edited}}, \hat{\mathbf{S}}),\end{aligned} \tag{5}$$

where $\mathcal{L}_{\text{MSE}}$ denotes the mean square error.

The only "learnable" variable is the editing vector δw^+ and all the neural networks are kept fixed. Note that as in EditGAN, there is a certain amount of ambiguity in how the segmentation modification is realized in the RGB output.

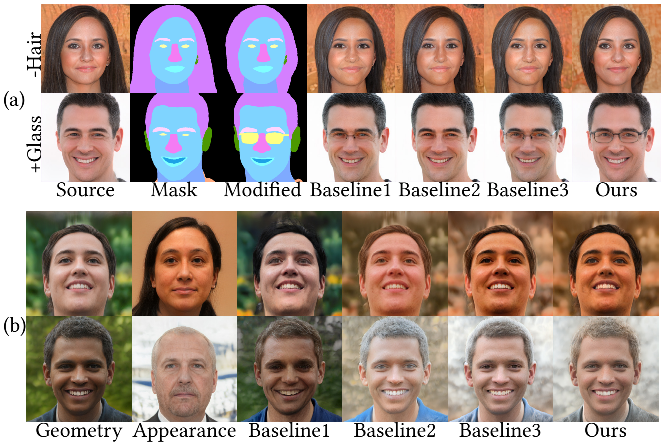

Figure 6: The results of the ablation study. (a) shows the ablation study for editing effects, while (b) shows the ablation study for style transfer.

4 EXPERIMENTS

In this section, we show our experimental setup and discuss the results of our experiments. Results from comparison with alternative methods, ablation study, and user study all show the effectiveness of our method and its superiority to the alternative approaches.

Implementation Details. Our decoder is implemented as a lightweight MLP with one hidden layer of 64 units. We use the Adam [Kingma and Ba 2015] optimizer with $\beta_1 = 0, \beta_2 = 0.99$ and the learning rate is fixed to 0.002 during fine-tuning. The fine-tuning takes 96 hours on 1 Tesla V100 GPU. NeRFFaceEditing is implemented in PyTorch [Paszke et al. 2019] and Jittor [Hu et al. 2020].

4.1 Results and Evaluations

In terms of rendering speed at inference time, our method achieves nearly real-time framerates at 512^2 resolution. On a single Tesla V100 GPU, we could render 20 images per second without tri-plane caching or 32 with tri-plane caching. Besides, we make comparisons by inverting real images in the wild with SofGAN [Chen et al. 2022a], which is also a view-consistent method.

Qualitative results. Our framework can generate semantic masks as well as realistic images from a certain pose. Thus, it can be used to edit the original facial volume through GAN inversion (described in Sec. 3.4), as shown in Fig. 5 (please see the supplementary materials for more results). For comparison, we perform editing on the real images as shown in Fig. 7(a). SofGAN achieves reasonable results at original poses, but disturbs the identity at other poses and fails to produce facial details (e.g., in the mouth and eyes) naturally. In contrast, our method outperforms SofGAN in both identity preservation and editing fidelity.

As for style transfer, our framework can disentangle the geometry and appearance of a facial volume and generate a new facial volume by swapping geometry and/or appearance. Thus, it can be used for 3D-aware style transfer, as shown in Fig. 4 (please refer to the supplementary materials for more results). For comparison, we perform style transfer between two real images. From Fig. 7(b), it

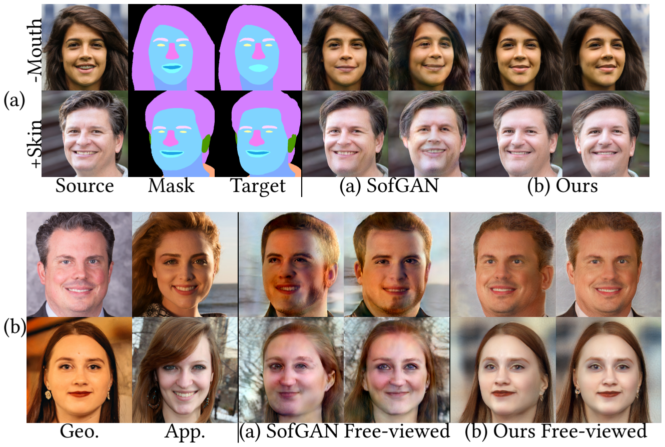

Figure 7: Comparisons with SofGAN [Chen et al. 2022a] in terms of editing (a) and style transfer (b). Original images courtesy of Thomas Rodenbücher, Brett Morrison, United Way of Central Ohio, Colin Brown, Krists Luhaers and Mckinnon de Kuyper.

is clear that during style transfer, SofGAN has obvious artifacts, especially for poses different from the original poses. In contrast, our method not only successfully transfers the style but also preserves the geometry accurately.

Quantitative evaluations. We measure image quality with Frechet Inception Distance (FID) [Heusel et al. 2017] on the FFHQ dataset [Karras et al. 2019]. Before fine-tuning, the FID of our implemented EG3D is 9.7, while after fine-tuning the FID increases to 16.0. Note that our fine-tuning needs less strict hardware requirements and shortened training time, at the cost of slightly degraded quality.

4.2 Ablation Study

We conduct ablation studies to justify the necessity of each component in our framework.

For the ablation study of editing, we show the edited results of "Baseline1", which uses a decoder to predict densities and semantic labels from the original tri-planes. The color features are reused from the original fixed decoder Φ. Thus, the geometry and appearance are not disentangled in this case. We also test "Baseline2", in which we use an off-the-shelf 2D segmentation module [Yu et al. 2018] to parse generated images into semantic masks for performing edits. Thus, the editing by optimization is performed on the fixed EG3D directly. Besides, we also show the results of "Baseline3", for which we train a decoder to predict semantic labels from the original tri-planes, while reusing densities and color features from the fixed original decoder Φ. Thus the semantic mask volume is not aligned with the facial volume. From Fig. 6(a), it is clear that without the disentanglement ("Baseline1"), with the off-the-shelf segmentation module ("Baseline2"), or without the alignment between geometry and semantic masks ("Baseline3"), the editing effects fail to be consistent with the modified semantic masks, such as the inaccurate hair length and incomplete glasses in the first and second rows.

For the ablation study of style transfer, we show the results by applying style-mixing (denoted as "Baseline1"), since the latent codes of high-resolution blocks roughly control the appearance (details can be found in the supplementary). Additionally, we show the results of the generator decoupled by our method but without $\mathcal{L}_{\text{Sim}}$ (denoted as "Baseline2"). We also give results with the histogram loss for a whole image instead of each label (denoted as "Baseline3"). From the part (b) of Fig. 6, it is clear that with the style-mixing ("Baseline1") or without $\mathcal{L}_{\text{Sim}}$ ("Baseline2"), the transferred style is inconsistent with the appearance reference. Furthermore, when the histogram color loss is calculated for the whole image, Baseline3 generally captures the style but is not accurate at details. For example, the color of background diffuses to the region of hair.

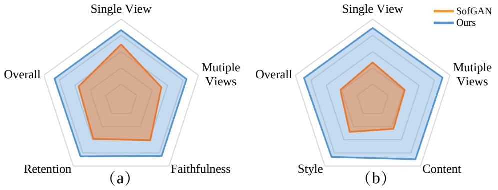

Figure 8: Radar plots of the average quality and faithfulness perception scores in terms of the similarity between the generated faces and the reference faces in appearance and geometry respectively, as well as the visual realism of the generated faces, in the single and multiple views. (a) The comparison of editing with two methods: SofGAN [Chen et al. 2022a] and ours. (b). the evaluation of style transfer with SofGAN and our method.

4.3 Perception Study

To evaluate the visual quality and the faithfulness of synthesized faces (i.e., the similarity to the geometry and appearance images), we conducted perception studies.

Specifically, we evaluate the performance of SofGAN and our method for editing and style transfer using two respective online questionnaires, in which each question is rated in a five-point Likert scale (1=strongly negative to 5=strongly positive). In the first questionnaire, we showed the original image and semantic mask, the modified semantic mask and results in the original view and multiple other views by the two methods, placed side by side in a random order to avoid bias. Each participant was asked to evaluate 15 examples according to five criteria: the visual quality of generated face images in the original view, the visual quality of generated face images in other views, faithfulness to the changed regions, retention to the unchanged regions, and overall effects. In total, 25 participants (7 female, 18 male, aged from 18 to 32, with normal vision) without any special experience participated in this study and we got 25 (participants) × 75 (questions) = 1875 subjective evaluations for each method. We draw a radar plot (see Fig. 8 (a)) in aspects of "Single View", "Multiple Views", "Faithfulness", "Retention", and "Overall" corresponding to the above-mentioned five criteria respectively. We found significant effects of our method for all five criteria based on scores in order: 4.32, 4.23, 4.23, 4.26 and 4.31 over 3.44, 2.60, 3.04, 2.92, 2.75 of SofGAN.

In the second questionnaire, we presented users with the geometry and the appearance reference images, and the results by SofGAN and ours in the original view and multiple other views.

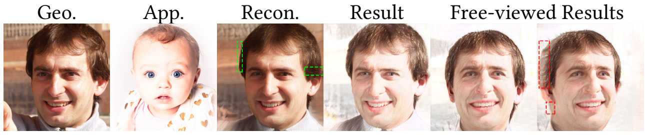

Figure 9: A failure case. The red rectangles show the artifacts such as shadows or on-the-background neck textures. The green rectangles show the lost details such as the hair fringe through PTI. Original images courtesy of pedronchi and Jane Gross.

Each participant was asked to evaluate 15 examples according to five criteria: the visual quality of synthesized face images in the original view, the visual quality of synthesized face images in other views, the maintenance of the geometry reference, the similarity to the appearance reference, and overall effects. In total, we got 25 (participants same to the first questionnaire) × 75 (questions) = 1875 subjective evaluations for each method. Fig. 8 (b) shows the statistics of these two methods, in which "Single View", "Multiple Views", "Content", "Style", and "Overall" correspond to the above-mentioned five criteria respectively. We get the values 4.45, 4.52, 4.48, 4.30 and 4.44 compared to 2.33, 2.07, 2.16, 2.40, 2.09 of SofGAN. It is clear that our method achieved a significant improvement over SofGAN.

We have done the analysis of one-way ANOVA tests and paired t-tests for the two questionnaires with $p < 0.001$ for all the tests, which confirmed our method significantly outperforms SofGAN. For more detail, please refer to the supplementary materials.

5 CONCLUSION, LIMITATIONS AND FUTURE WORK

As the training for high-quality 3D-aware GANs is more and more time-consuming and resource-hungry, approaches for intuitive controlling of geometry and appearance based on pretrained 3D GANs should receive more attention. We believe that our method, a structured disentanglement framework based on the tri-plane representation, is a step forward in this direction, as we demonstrated our method can effectively edit the geometry through semantic masks and perform style transfer with fine-tuning while preserving the high speed and visual quality of the backbone. In the future, we are interested in implementing our framework based on the official EG3D and evaluating the impact of such a change.

One limitation of this work is that due to the generation quality and inversion approach of our model, our method has artifacts such as shadows and some "billboards" effects at relatively large angles and fails to reach faithful and stereoscopic reconstruction (see Fig. 9). We speculate that they may come from our specific training process. Besides, in Fig. 9, rendered images are sometimes not realistic enough when performing style transfer. We attribute it to light and shadow, which our method cannot handle well. As future work, it would be useful to explore disentanglement of other attributes such as lighting to make our method more general. Besides, semantic masks have inevitable ambiguities while performing optimization on the single view, such as failing to control the gender or holding unseen parts consistent.

Our method is capable of converting a real portrait image to its 3D avatar and even possibly a talking head through editing guided by semantic masks. Moreover, we can change its appearance while keeping its geometry unchanged. The editing may disturb the gender, and the style transfer may disturb the ethics of the original portrait. Therefore misusing our system potentially poses a societal threat, including ethics issues and fooling facial recognition systems. We do not condone using our work with the intent of spreading misinformation or tarnishing reputation. Thus, one should be careful to deploy this technology. However, in case of misusing, existing methods (e.g., [Dang et al. 2020]) for detecting fake faces might alleviate this concern.

ACKNOWLEDGMENTS

Thanks to Ms. Xiaohong Wang for her valuable comments and suggestions on the use of OneITLab. This work was supported by grants from the National Natural Science Foundation of China (No. 62061136007, No. 62102403, and No. 61872440), Science and Technology Service Network Initiative of the Chinese Academy of Sciences (No. KFJ-STS-QYZD-2021-11-001), the Beijing Municipal Natural Science Foundation for Distinguished Young Scholars (No. JQ21013), ChinaPostdoctoral Science Foundation (No. 2022M713205), the Youth Innovation Promotion Association CAS. Hongbo Fu was supported by the Research Grants Council of HKSAR (No. 11212119), and the Centre for Applied Computing and Interactive Media (ACIM) of School of Creative Media, City University of Hong Kong.

REFERENCES

Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. 2021. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (TOG)* 40, 3 (2021), 21:1–21:21.

Mahmoud Afifi, Marcus A Brubaker, and Michael S Brown. 2021. Histogan: Controlling colors of gan-generated and real images via color histograms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 7941–7950.

ShahRukh Athar, Zhixin Shu, and Dimitris Samaras. 2021. Flame-in-nerf: Neural control of radiance fields for free view face animation. *arXiv preprint arXiv:2108.04913* (2021).

Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, et al. 2021a. Efficient geometry-aware 3D generative adversarial networks. *arXiv preprint arXiv:2112.07945* (2021).

Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. 2021b. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 5799–5809.

Anpei Chen, Ruiyang Liu, Ling Xie, Zhang Chen, Hao Su, and Jingyi Yu. 2022a. Sofgan: A portrait image generator with dynamic styling. *ACM Transactions on Graphics (TOG)* 41, 1 (2022), 1:1–1:26.

Shu-Yu Chen, Feng-Lin Liu, Yu-Kun Lai, Paul L Rosin, Chunpeng Li, Hongbo Fu, and Lin Gao. 2021. DeepFaceEditing: deep face generation and editing with disentangled geometry and appearance control. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 90:1–90:15.

Shu-Yu Chen, Jia-Qi Zhang, You-You Zhao, Paul L. Rosin, Yu-Kun Lai, and Lin Gao. 2022c. A review of image and video colorization: From analogies to deep learning. *Visual Informatics* (2022). https://doi.org/10.1016/j.visinf.2022.05.003

Yuedong Chen, Qianyi Wu, Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. 2022b. Sem2NeRF: Converting Single-View Semantic Masks to Neural Radiance Fields. *arXiv preprint arXiv:2203.10821* (2022).

Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil Jain. 2020. On the Detection of Digital Face Manipulation. In *In Proceeding of IEEE Computer Vision and Pattern Recognition.* Seattle, WA.

Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. 2022. GRAM: Generative Radiance Manifolds for 3D-Aware Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2021. Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 8649–8658.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2016. Image Style Transfer Using Convolutional Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2414–2423.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. 2021. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985* (2021).

Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. 2021. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5784–5794.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. 2021. Headnerf: A real-time nerf-based parametric head model. *arXiv preprint arXiv:2112.05637* (2021).

Shi-Min Hu, Dun Liang, Guo-Ye Yang, Guo-Wei Yang, and Wen-Yang Zhou. 2020. Jittor: a novel deep learning framework with meta-operators and unified graph execution. *Science China Information Sciences* 63, 222103 (2020), 1–21.

Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1501–1510.

Wonbong Jang and Lourdes Agapito. 2021. Codenerf: Disentangled neural radiance fields for object categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12949–12958.

Kyungmin Jo, Gyumin Shim, Sanghun Jung, Soyoung Yang, and Jaegul Choo. 2021. CG-NeRF: Conditional Generative Neural Radiance Fields. *arXiv preprint arXiv:2112.03517* (2021).

Kacper Kania, Kwang Moo Yi, Marek Kowalski, Tomasz Trzciński, and Andrea Tagliasacchi. 2022. CoNeRF: Controllable Neural Radiance Fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4401–4410.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8110–8119.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*.

Chuan Li and Michael Wand. 2016. Combining markov random fields and convolutional neural networks for image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2479–2486.

Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. 2017a. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (TOG)* 36, 6 (2017), 194:1–194:17.

Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. 2017b. Demystifying Neural Style Transfer. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. 2230–2236.

Connor Z Lin, David B Lindell, Eric R Chan, and Gordon Wetzstein. 2022. 3D GAN Inversion for Controllable Portrait Image Animation. *arXiv preprint arXiv:2203.13441* (2022).

Huan Ling, Karsten Kreis, Daiqing Li, Seung Wook Kim, Antonio Torralba, and Sanja Fidler. 2021. EditGAN: High-Precision Semantic Image Editing. *Advances in Neural Information Processing Systems* 34 (2021).

Steven Liu, Xiuming Zhang, Zhoutong Zhang, Richard Zhang, Jun-Yan Zhu, and Bryan Russell. 2021. Editing conditional radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5773–5783.

N. Max. 1995. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics* 1, 2 (1995), 99–108.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European Conference on Computer Vision*. 405–421.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *arXiv:2201.05989* (Jan. 2022).

Michael Niemeyer and Andreas Geiger. 2021. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11453–11464.

Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. 2021. StyleSDF: High-Resolution 3D-Consistent Image and Geometry Generation. *arXiv e-prints* (2021), arXiv–2112.

Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan Goldman, Steven Seitz, and Ricardo Martin-Brualla. 2020. Nerfies: Deformable Neural Radiance Fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5845–5854.

Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. 2021. HyperNeRF: a higher-dimensional representation for topologically varying neural radiance fields. *ACM Transactions on Graphics (TOG)* 40, 6 (2021), 238:1–238:12.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. 2021. Pivotal tuning for latent-based editing of real images. *arXiv preprint arXiv:2106.05744* (2021).

Sara Fridovich-Keil and Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. 2022. Plenoxels: Radiance Fields without Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. 2020. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems* 33 (2020), 20154–20166.

Jingxiang Sun, Xuan Wang, Yichun Shi, Lizhen Wang, Jue Wang, and Yebin Liu. 2022. IDE-3D: Interactive Disentangled Editing for High-Resolution 3D-aware Portrait Synthesis. *arXiv preprint arXiv:2205.15517* (2022).

Jingxiang Sun, Xuan Wang, Yong Zhang, Xiaoyu Li, Qi Zhang, Yebin Liu, and Jue Wang. 2021. FENeRF: Face Editing in Neural Radiance Fields. *arXiv preprint arXiv:2111.15490* (2021).

Luan Tran and Xiaoming Liu. 2018. Nonlinear 3d face morphable model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7346–7355.

Ziyan Wang, Timur Bagautdinov, Stephen Lombardi, Tomas Simon, Jason Saragih, Jessica Hodgins, and Michael Zollhofer. 2021. Learning Compositional Radiance Fields of Dynamic Human Heads. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5704–5713.

Yinghao Xu, Sida Peng, Ceyuan Yang, Yujun Shen, and Bolei Zhou. 2021. 3D-aware Image Synthesis via Learning Structural and Textural Representations. (2021).

Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. 2021. PlenOctrees for Real-time Rendering of Neural Radiance Fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. 2018. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European Conference on Computer Vision*. 325–341.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 586–595.

Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C. Bühler, Xu Chen, Michael J. Black, and Otmar Hilliges. 2022. I M Avatar: Implicit Morphable Head Avatars from Videos. In *Computer Vision and Pattern Recognition (CVPR)*.

Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. 2021. Cips-3d: A 3d-aware generator of gans based on conditionally-independent pixel synthesis. *arXiv preprint arXiv:2110.09788* (2021).

Yiyu Zhuang, Hao Zhu, Xusen Sun, and Xun Cao. 2021. MoFaNeRF: Morphable Facial Neural Radiance Field. *arXiv preprint arXiv:2112.02308* (2021).