SemFaceEdit: Semantic Face Editing on Generative Radiance Manifolds

Shashikant Verma[0000−0002−9862−1379] and Shanmuganathan Raman∗[0000−0003−2718−7891]

CVIG Lab, Indian Institute of Technology Gandhinagar, India
{shashikant.verma, shanmuga}@iitgn.ac.in

Abstract. Despite multiple view consistency offered by 3D-aware GAN techniques, the resulting images often lack the capacity for localized editing. In response, generative radiance manifolds emerge as an efficient approach for constrained point sampling within volumes, effectively reducing computational demands and enabling the learning of fine details. This work introduces SemFaceEdit, a novel method that streamlines the appearance and geometric editing process by generating semantic fields on generative radiance manifolds. Utilizing latent codes, our method effectively disentangles the geometry and appearance associated with different facial semantics within the generated image. In contrast to existing methods that can change the appearance of the entire radiance field, our method enables the precise editing of particular facial semantics while preserving the integrity of other regions. Our network comprises two key modules: the Geometry module, which generates semantic radiance and occupancy fields, and the Appearance module, which is responsible for predicting RGB radiance. We jointly train both modules in adversarial settings to learn semantic-aware geometry and appearance descriptors. The appearance descriptors are then conditioned on their respective semantic latent codes by the Appearance Module, facilitating disentanglement and enhanced control. Our experiments highlight SemFaceEdit's superior performance in semantic field-based editing, particularly in achieving improved radiance field disentanglement.

Keywords: Neural Radiance Fields · Neural Rendering · 3D-aware GANs

1 Introduction

In recent years, there has been significant interest in vision and graphics in generating visually captivating and photo-realistic images. Notably, 3D aware Generative Adversarial Networks incorporating adversarial learning [11] and Neural Radiance Fields [25] have made remarkable progress in producing multiple-view images that closely resemble real photographs. Despite advancements, 3D aware Generative Adversarial Networks (GANs) [28, 6, 12] still struggle to match the resolution of state-of-the-art 2D GAN models. This limitation arises from the

---

high computational demands of learning volumetric representations. Recently proposed methods [5, 8] try to efficiently learn radiance fields by a tri-plane representation and learnable radiance manifolds, respectively. Nevertheless, these methods lack control over learned geometry, thus posing challenges in decoupling appearance from geometry.

To decouple facial geometric and appearance attributes, FENeRF [35] adopts π-GAN [6] and learns facial semantic volume aligned with geometry. The learned facial semantics act as an intermediary, offering control over geometry and appearance. However, implementing local editing with FENeRF necessitates time-consuming and resource-intensive retraining. Addressing this limitation, [15] proposes an AdaIN-based Controllable Appearance Module (CAM) and a geometry decoder. This approach allows for independent control of appearance from geometry. Unlike FENeRF, this approach requires training a geometry decoder on learned tri-plane representations from [5] to obtain the semantic volume for editing. Consequently, the CAM module in [15] grants control over the appearance of the entire generated portrait. While introducing the geometry decoder branch facilitates disentangled face editing, it does not provide semantic control over appearance.

In this work, we introduce SemFaceEdit, building upon efficient Generative Radiance Manifolds (GRAM) [8]. GRAM employs 2D manifolds to constrain point sampling and radiance field learning, jointly trained with GAN. This constrained approach reduces the computational workload and enables fine detail learning by confining point sampling and radiance learning within a reduced space. In contrast to GRAM, our approach learns both semantically disentangled geometry and appearance in a unified manner. We introduce two modules: the Geometry Module and the Appearance Module. The Geometry Module predicts the semantic radiance field within a volume, which the Appearance Module subsequently uses to condition the appearance semantically. Unlike existing methods that can entirely change the appearance or geometry of a radiance field, our proposed approach enables semantic-specific editing while preserving the integrity of other semantic information in terms of geometry or appearance. Our method facilitates changes in geometry and appearance within a single photograph and enables seamless transfer of hairstyle or other semantic facial attributes from one source reference to another.

The main contributions of this work are summarized as follows.

- We present a novel framework based on generative radiance manifolds, allowing for semantic control over the geometry and appearance by manipulating latent spaces specific to each semantic attribute.
- We design a differentiable Semantic Volume masking layer, which learns to segregate each point in radiance volume into semantic groups. This layer enables targeted appearance transfer concerning aspects like colours and hues and selective geometry transfer for various semantic regions.
- Our method excels in geometry and appearance control, allowing for the seamless transfer of geometry or appearance associated with one semantic label in a radiance field to another.
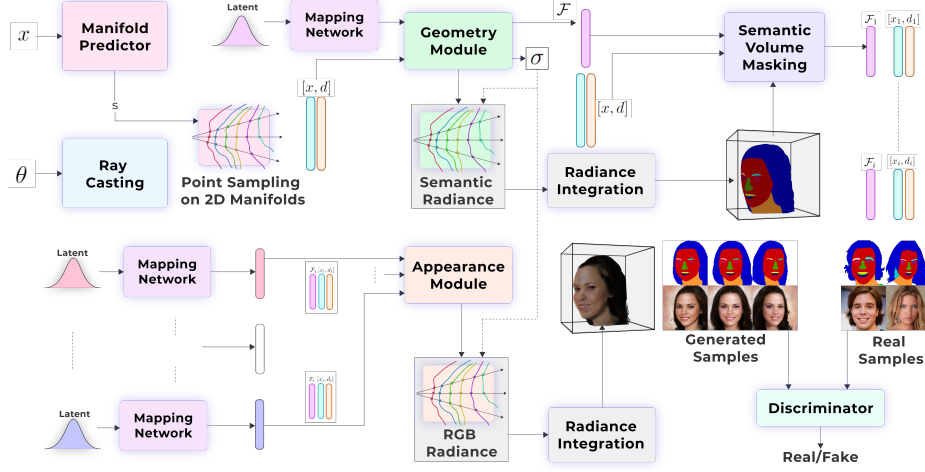
Fig. 1. An overview of our proposed framework. We sample points in volume by determining intersections of casted rays with isosurfaces predicted by Manifold Predictor [8]. Subsequently, the Geometry Module conditions these points using a latent vector sampled from a Gaussian distribution, resulting in diverse predictions for occupancy (σ), semantic radiance, and Appearance Descriptor $\mathcal{F}$. To segregate points and their appearance descriptors based on semantic classes, we employ the Semantic Volume Masking layer, which relies on semantic radiance information. The Appearance Module then utilizes different latent codes to condition each set of points and predicts RGB Radiance. Selectively conditioning Appearance Descriptors $\mathcal{F}$ based on semantics enables the Appearance Module to independently modify a specific semantic appearance.

2   Related Works

Neural Implicit Representations. Neural Implicit functions have been widely applied to various vision problems such as novel view synthesis [36, 32], and image manipulations [1, 40]. Furthermore, they have gained significant traction in diverse 3D-related tasks, including 3D scene reconstruction [25], and occupancy field or signed distance function estimation [28, 24]. Recent works have explored alternative representations of implicit functions, such as octrees and tri-planes [5, 37], for faster inference and improved expressive power. Additionally, GRAM [8] introduces an effective point sampling method and radiance field learning on 2D manifolds, enabling fine detailed learning with reduced computational complexity. This paper emphasizes learning semantic information on 2D manifolds through a geometry module rather than directly learning about the image space. However, learning the semantic field alone encounters challenges due to the absence of depth information in the semantic mask data. To address this, the learning of the semantic field is further supervised by an additional appearance module. Unlike [35] and [15], our appearance module uses the semantic infor-
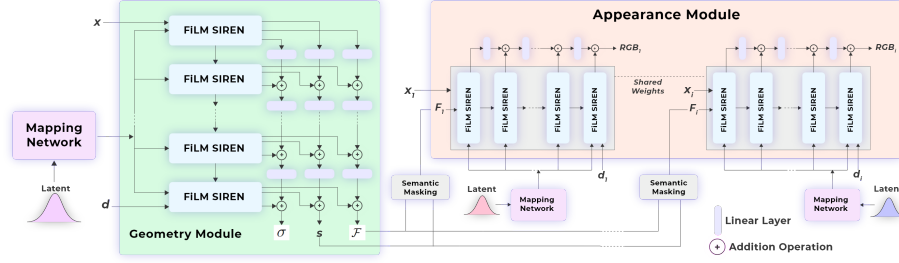
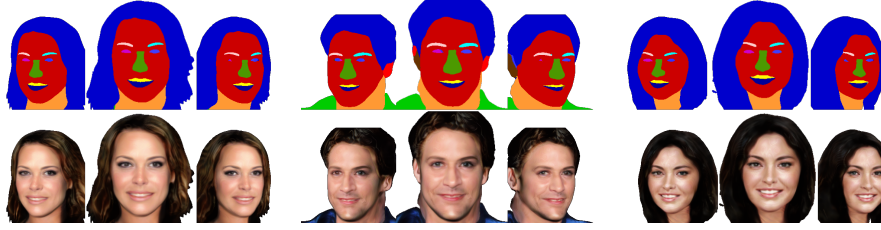**Fig. 2.** The network architecture of Geometry Module and Appearance Module.

Fig. 3. Renderings of Semantic-radiance and RGB-radiance on image space generated by our approach by random latent code $\mathbf{z} \in \mathrm{R}^d$ and $\mathbf{z}_i \in \mathrm{R}^d$.

mation from the geometry module to control the appearance of each semantic point on these manifolds in the volume.

3D-aware Neural Face Image Synthesis. The combination of Generative Adversarial Networks (GANs), as originally proposed [11], and neural implicit radiance fields (NeRF) [25], has been widely explored for the generation of 3D-consistent faces. Despite utilizing 3D aware features and 2D CNN-based renderers to generate synthetic images, several methods such as [12, 22, 26, 6] face challenges in accurately representing geometry, resulting in lower quality representations. Recent studies [2, 8, 5, 28] investigate novel representations and models, enabling the synthesis of geometrically consistent fine details.

Editing Faces with 2D GANs and DDPMs. Generative adversarial networks (GANs), and more recently, Denoising Diffusion Probabilistic Models (DDPMs) have widely been applied for generating realistic images, with face editing as a prominent area of research in this context. Various methods have been proposed for face editing using semantic maps. These approaches leverage conditional 2D GANs [29, 42, 7, 20], and DDPMs [14, 17, 33] which are conditioned on semantic masks. Unlike explicit field/occupancy-based neural scene representations, SofGAN [7] utilizes implicit fields for semantic occupancy generation. However, SofGAN relies on semantically labelled 3D scans for training and involves a separate process to synthesize images through 2D projections of semantic labels. Diffusion-Rig [9] first learn facial priors by fitting a morphable face model for conditioning diffusion model to generate view-consistent 2D facial images. In contrast, our proposed framework directly modifies appearance in the

generated implicit neural representation without the need for 2D projections or facial priors, and is trained on a collection of 2D images instead of 3D scans.

Editing Faces in Neural Radiance Fields. To enable control over the geometry and appearance of synthesized images by 3D-aware GANs, subsequent works employ the embedding of explicit control into the generation process. Certain methods, including [3, 43, 39], condition radiance fields with information from 3D Morphable models (3DMM) [21, 30] to exert control over the generated fields. This conditioning is achieved through embeddings of latent spaces representing pose, shape, and expression parameters of the 3DMM models. FENeRF [35] utilizes semantic masks for geometry control and a separate latent vector for appearance control in synthesized images. CG-NeRF [16] introduces soft conditions, including sketches as external inputs, to influence radiance field generation. Following [5], IDE-3D [34] enables editing on tri-plane representations while NeRFFaceEditing [15] disentangles geometry and appearance using pre-trained tri-plane representations, eliminating the need for retraining.

While the aforementioned methods provide control over the geometry and appearance of synthesized images, they fall short in offering semantic-specific control. Our proposed method provides seamless semantic-specific control over geometry and appearance attributes in facial images, as shown in Figure 6. This allows for effortlessly transferring specific facial attributes to other images, including overall geometry and appearance.

3 Methodology

Our primary objective is to learn semantic information, occupancy, and colour attributes for points sampled on 2D Manifolds. The Manifold Predictor model predicts these manifolds [8]. To separate the representation of geometry and texture during the image generation process, we introduce two distinct modules: the Geometry module and the Appearance module, as illustrated in Figure 1. The Geometry module employs a shape latent code for each point on the manifolds, influencing occupancy and semantic information. Simultaneously, it predicts high-dimensional Appearance Descriptors $\mathcal{F}$ for each point. These descriptors are segregated based on semantic information using the Semantic Volume Masking Layer. The Appearance module then conditions points belonging to a specific semantic with different appearance latent codes and predicts colour information. The entire method is trained in two stages using adversarial learning. Initially, we train the entire network end-to-end, simultaneously learning weights for the Geometry and Appearance modules. Later, we freeze the weights of the Geometry module and fine-tune the Appearance module to achieve high-quality volumetric renderings of generated images.

3.1 Network Architecture

Our generator consists of the Geometry Module and the Appearance Module, as illustrated in Figure 1. Both modules are implemented as Multi-Layer
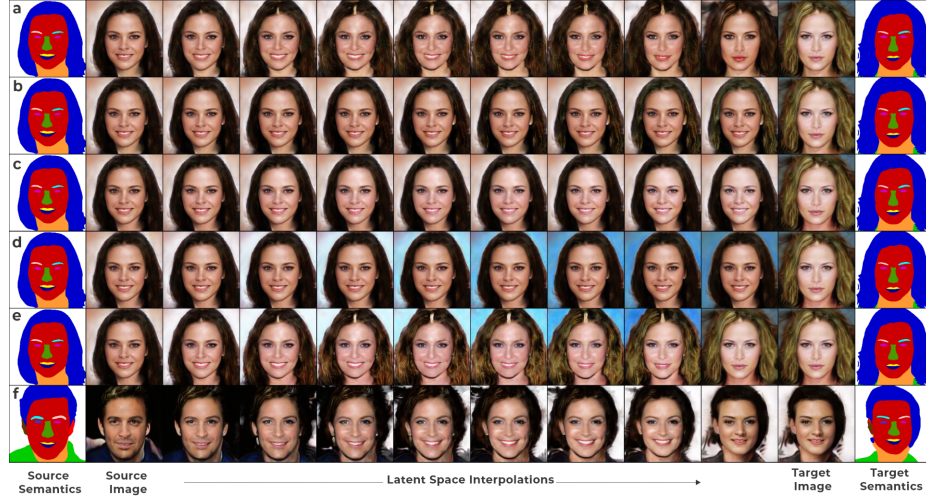
Fig. 4. Latent space disentanglement and interpolations. In (a), linear interpolation is performed on the geometric latent space z while keeping all appearance latent z_i fixed. (b-d) demonstrate changes in appearance latent variables corresponding to hair, face, and background semantics. All latent codes are linearly interpolated for (e) and (f). Additionally, (f) showcases the change in the appearance of garment semantics from the source to the target.

Perceptrons (MLPs) to produce occupancy σ, semantic radiance s, and colour $c = (r, g, b)$ values for a given point $x \in \mathrm{R}^3$ with view direction $d \in \mathrm{R}^3$.

$$\Psi_g : (\mathbf{z}, \mathbf{x}, \mathbf{d}) \mapsto (\sigma, \mathbf{s}), \; \Psi_a : (\{\mathbf{z_i}\}_{i=0}^{l}, \mathbf{x}, \mathbf{d}, \mathbf{s}) \mapsto (\mathbf{c}) \tag{1}$$

where, $\mathbf{z} \in \mathrm{R}^d \sim p_z$ represents the geometric latent code and $\mathbf{z_i} \in \mathrm{R}^d \sim p_z$ represents the appearance latent code for each semantic labels l.

Manifold Predictor. We sample N points along each ray cast in the direction $\mathbf{d}$, given a camera pose $\theta \in \mathrm{R}^3 \sim p_\theta$. The Manifold Predictor $\mathcal{M}$ is implemented as a scalar field function using light-weight MLPs, predicting a scalar value s for the sampled points. From the predicted scalar field, we extract K iso-surfaces with different levels l_i and perform final point sampling by finding the nearest point of intersection on these iso-surfaces, following [8].

Geometry Module. Given a latent code z, the Geometry module Ψ_g generates the radiance for each point on the predicted 2D manifolds in volume. The architecture of Ψ_g consists of FiLM SIREN backbone, inspired from [6] with some modifications, as presented in Figure 2. Instead of directly synthesizing images as in [8], our geometric module first learns to predict occupancy σ and semantic radiance field, $\mathbf{s}$. Our primary goal is to disentangle geometry from appearance by learning geometry independently of image synthesis. However, attempting to learn semantic radiance solely using semantic masks in an adversarial setting is bound to fail, as the perceptual appearance of a semantic mask lacks depth

information present in images due to the absence of shading effects caused by varying shades of light.

To address this limitation, we additionally predict high-dimensional feature descriptors $\mathcal{F}$ for each point, which is then conditioned by the Appearance module for image synthesis. By backpropagating gradients through $\mathcal{F}$ from the appearance module into the geometry module, we ensure the correct learning of the occupancy field σ. This approach helps in better disentangling geometry and appearance, leading to improved results in our framework.

Semantic Volume Masking Layer. Given Semantic radiance field $\mathbf{s}$ and occupancy field σ, we perform volumetric rendering [27, 41] on all 2D manifolds. When a ray r is cast and intersects K surface manifolds at points x_j sorted by proximity, we can describe the semantics of a point on the k-th manifold out of the total K manifold:

$$\mathrm{S}_k(r) = \arg\max \sum_{j=k}^{K} T(\mathrm{x}_j)\sigma(\mathrm{x}_j)\mathrm{s}(\mathrm{x}_j, \mathrm{d}), \ \ \textbf{for} \ \ k \leq K \tag{2}$$

Here, $T(x_j) = \prod_{i<j; i \geq k}(1 - \sigma(\mathrm{x}_j))$, and Equation 2 allows us to classify each sampled point to the semantic class it belongs to. After obtaining the semantic labels, we segregate the points and their corresponding appearance descriptors $\mathcal{F}$ based on their specific semantic categories. Using the Appearance module, this segregation allows us to condition them on respective semantic-specific latent codes.

Let there be n semantic classes, and the Semantic Volume Masking layer segregates all sampled points in the volume into n collections, denoted by x_i along with their corresponding casted direction d_i and appearance descriptor $\mathcal{F}_i$. These collections are created such that every point $x \in \mathrm{x}_i$ belongs to a specific semantic category i, defined among the n available semantic classes.

Appearance Module. The architecture of our appearance module closely resembles that of the Geometry Module, as depicted in Figure 2. However, it is distinguished by the presence of n mapping networks, each responsible for conditioning the i-th collection obtained from the Volume Masking layer. These mapping networks are utilized to condition the Appearance Descriptors $\mathcal{F}_i$, which then pass through a SIREN-based neural network for RGB radiance prediction. As a result, our appearance module Ψ_a generates a radiance field c, where the i-th latent code controls the appearance of the i-th semantic among all n semantic classes. Importantly, it should be noted that while different mapping networks are used for conditioning the collections, the weights of the SIREN backbone in the appearance module remain consistent across all collections. In the Results Section, we present exhaustive ablation studies conducted on the Appearance Module to identify the optimal network architecture. The Appearance Module takes segregated Appearance Descriptors from the geometry module and a latent code for each semantic as input, producing the final RGB radiance. Subsequently, we perform volumetric radiance integration with the occupancy field σ, obtained from the Geometry Module, following a similar approach as in [27, 8], to obtain the final image.

Fig. 5. Appearance Latent Codes (z_i) influence on Geometric Shapes. The image (i,j) in grid (a),(b), and (c) are generated with same z_i's, highlighting resemblance in appearance across hair, background, and facial regions with (i,j)th image in each grid.

4   Experiments and Results

4.1   Training

We train SemFaceEdit on the CelebAMask-HQ [19] dataset. The dataset contains 30K high-resolution face images and annotated masks in $N = 19$ classes. In our experiments, we club N classes in 4 semantics, denoted as n: hairs, face (containing eyes, nose, mouth, neck, and ears), garment, and background. We train our network in two stages in an adversarial learning setting, using non-saturating GAN loss with R1 regularization [23]. During training, we randomly sample geometric latent code z, n appearance latent codes z_i and pose θ from Gaussian distributions $p_z^g, p_{z_i}^a$, and p_θ, respectively. The generator, consisting of Ψ_g and Ψ_a, produces semantic s, occupancy σ, and RGB c, radiance fields. In the first stage, we jointly train Ψ_g and Ψ_a with two discriminators, Semantic map discriminator D_s and Facial image discriminator D_c. We provide further details on data preprocessing, discriminator architecture and loss functions in *Suppl. material*. In Figure 3, We present multi-view images generated from randomly sampled latent codes after training the model.

Implementation Details. Our experiments utilize the Adam Optimizer [18] with learning rates of 2×10^{-5} for the generator and 2×10^{-4} for the discriminator. Utilizing 2 NVIDIA GeForce RTX 4090 GPUs, each with 24 GB of memory, the model is trained for 120K iterations (3 epochs) during the first stage, followed by 30K iterations (1 epoch) in the second stage. To generate neural rendering of radiance field on to image space $\mathcal{I} \in \mathrm{R}^{W \times H \times 3}$, we cast $W \times H$ rays from the camera and sample K number of points along each ray, where $K = 24$ is number of iso-surfaces on manifolds. Specifically, we sample radiance values of 3.14 million ($256 \times 256 \times 24 \times 2$) points in total from semantic-radiance field s and rgb-radiance field c. To overcome computational challenges, we utilize
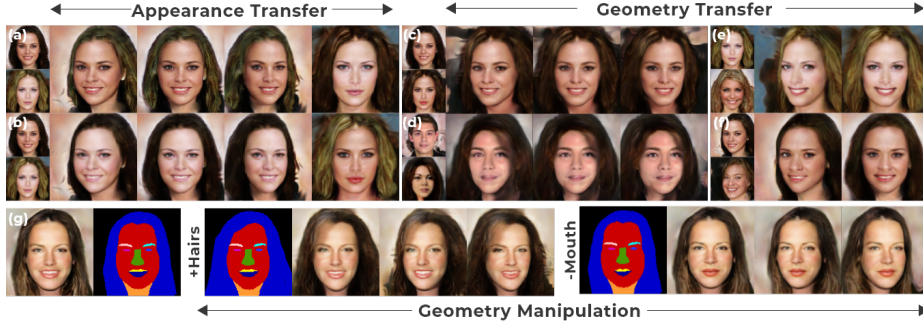
Fig. 6. (a-f) displays two smaller images, with the top image representing the source attribute and the bottom image representing the target attribute of appearance/geometry. **Appearance Transfer**: In (a), we present three different views of the source image showcasing the transferred hairstyle appearance from the target to the source image. Additionally, we present a single view of a converse transfer of appearance, i.e. from the source to the target image. Similarly, in (b), we showcase the transferred facial appearance. **Geometry Transfer**: (c) and (d) demonstrate the transfer of target hair geometry to the source image. In contrast, (e) and (f) showcase the smile and nose geometry transfer from the target to the source image. Note that in (c) and (d), alterations in attributes at regions other than hair semantics are minimal. In contrast, the transferred hair adopts the appearance style of the target image through manipulation of the latent code governing hair semantics. The focus is solely on transferring geometry for finer semantics like the nose or smile in (e, f). Therefore, we constrain appearance transfer by maintaining the unchanged state of appearance latents. **Geometry Manipulation**: (g) Shows the effect of manipulation in the semantic mask by expanding the hair region semantics and the effect of shrinking at the mouth region.

PyTorch's Automatic Mixed Precision (AMP) to reduce memory cost. During inference, our approach generates 256^2 and 128^2 image in 0.86 and 0.15 seconds with peak GPU memory usage of approx. 15GB and 6GB, respectively.

4.2   Editing Faces in Radiance Fields.

To edit an input image obtained either from I $\sim p_{rim}$ (real images) or synthetically generated using Generators Ψ_g and Ψ_a, we perform image inversion into the $\mathcal{W}$ space, represented as w, utilizing pivotal tuning inversion [31]. To enable image editing based on semantic mask manipulations, we require the original mask S and the user-edited mask S'. For real face images, we utilize the method proposed by [38] to obtain semantic mask S. On the other hand, for synthetic images generated through our pipeline, the Geometry module Ψ_g naturally provides us with the required semantic mask. Our experiments utilize trained mapping networks to obtain the average frequency and phase shifts over 10k sampled latent codes for geometry z and appearance z_i. For brevity, we denote these averaged frequencies and phase shifts as a latent code $w \in \mathcal{W}$ and
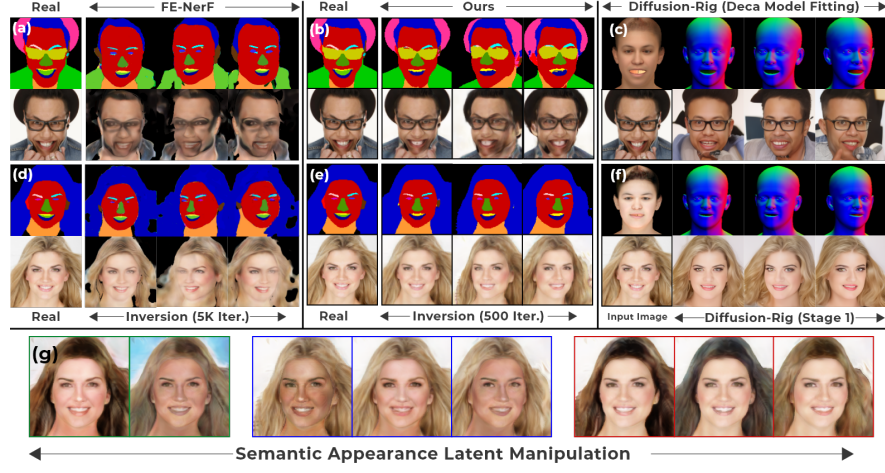
Fig. 7. Comparison of pivotal inversion results between our proposed approach (b,e) and FE-NerF [35] (a,d). Our method achieves rapid convergence within 500 iterations, contrasting with FE-NerF's 5K iterations. Additionally, we compare with Diffusion-Rig [9] (c,f), which first estimate facial prior by fitting a Deca model [10] on given image. Subsequently it employs a diffusion-based approach to generate view consistent facial images. (g) demonstrates semantic appearance variation achieved by modifying latent codes specific to distinct semantic regions $\mathbf{z_i}$ while maintaining geometry latent $\mathbf{z}$ of (d). Green outline signifies full appearance alteration; blue outlines indicate facial appearance changes and red outlines represent hair appearance modifications. Real images for inversion are taken from CelebAMask-HQ Dataset [19].

$w_i \in \mathcal{W}_i$ for geometry and appearance, respectively. To achieve editing, we optimize an editing offset vector $\delta w^+ \in \mathcal{W}$ such that the generated semantic mask S' from Ψ_g approximates the original mask S. During optimization, we keep w_i for Ψ_a fixed while optimizing for the geometric latent δw^+. We directly employ the corresponding z_i's for the appearance of synthetically generated images. For real images, we further optimize for δw_i^+ to make the appearance resemble the given image I. Figure 6 showcases editing in geometry based on semantic masks, and Figure 7(b,d) demonstrates inversion and latent code manipulations on real images from CelebAMask-HQ Dataset.

Geometry and Appearance Transfer. We can directly synthesize appearance manipulations targeting a specific semantic $k \in n$ using Ψ_a. Unlike [15], Ψ_a controls the change in appearance of regions corresponding only to the k-th semantic. We illustrate these results qualitatively in Figures 5 and 4. However, as discussed previously, image inversion for Geometry transfer into the $\mathcal{W}$ space is required. Given images I$_1$ and I$_2$ with semantics S_1 and S_2, respectively, let (w^1, w^2) represent geometric latents and (w_i^1, w_i^2) be appearance latent codes. We want to transfer the semantic region $k \in n$, represented by $\mathcal{M}_k$, from image I$_2$ to I$_1$. For appearance transfer, we directly interchange the k-th appearance semantic latent code in w_i^1 with w_k^2, and then optimize for δw^+, such that $\Psi_g(w^1+$
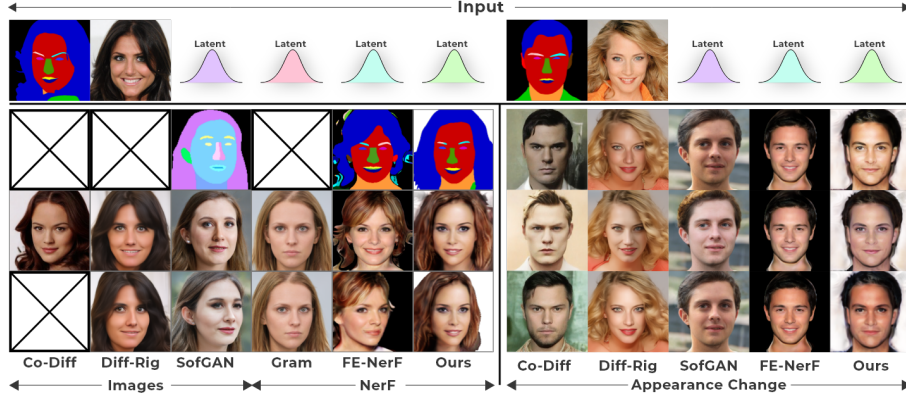
Fig. 8. Qualitative comparison of SemFaceEdit with Co-Diff [14], Diff-Rig [9], SofGAN [7], GRAM [8], and FE-NerF [35]. ⊠ represents that the respective approach don't support or yield the required result. Input mask in [14] and image in [9] are taken from CelebAMask-HQ Dataset[19].

$\delta w^+, \theta)$ approximates the target semantic $S_1 \odot (1 - \mathcal{M}_k) + S_2 \odot \mathcal{M}_k$, and $\Psi_a(w_i^1, \theta)$ approximates the target image $I_1 \odot (1 - \mathcal{M}_k) + I_2 \odot \mathcal{M}_k$. Though w_i^1 for appearance is fixed, inversion requires a backward pass through the Ψ_g and Ψ_a to update δw^+. This is because Ψ_g depends on Ψ_a for the correct estimation of appearance feature descriptors $\mathcal{F}$ and occupancy σ. In Figure 6 (c-g), we show the geometry transfer from target to source of various semantics. More details can be found in the *Suppl. Material*.

4.3   Results

We conduct a comparative analysis of SemFaceEdit against five state-of-the-art techniques: Co-Diff [14], Diff-Rig [9], GRAM [8], FE-NerF [35], and SofGAN [7]. GRAM and FE-NerF are oriented towards generating 3D-aware images, while SofGAN employs a 2D projection of the 3D semantic occupancy field for dynamic styling through a 2D GAN methodology. In contrast to SofGAN, Diff-Rig [9] employs a diffusion-based approach for view-consistent 2D image generation. This method lacks semantic-guided editing capability, instead utilises expression basis and priors from a morphable face model for appearance changes. We also include Co-Diff [14], a 2D facial image generation approach, which generates realistic facial images given a semantic map but lacks multiple view generation. **Qualitative comparison**. Our framework estimates semantic radiance in addition to the occupancy field and RGB radiance. In Figure 8 we showcase the generated semantic maps and multi-view images produced by our method compared to the aforementioned approaches. Additionally, on the right, we display the appearance manipulation achieved by our method in comparison with the other techniques. Co-Diff takes a semantic map as input and generates a facial image that fits the provided map, but it cannot generate multi-view images.

Table 1. Comparing FID and KID between 10K generated and 10K real CelebA-MaskHQ [19] Dataset images with GRAM [8], FE-NerF [35], SofGAN [7], Diffusion-Rig [9], and Co-Diff [14]. Lower values (↓) indicate better performance. Superfixes *
and † indicates that methods do not generate neural rgb-radiance fields but employ
2-dimensional deep neural networks (GANs/DDPMs) to synthesize multi-view consistent images (*) or only a single view image (†). Abbreviations used: GAC - Global
Appearance Control, LAC - Local Appearance Control, and SGE - Semantic Guided
Editing. Note that Diffusion-Rig do not support SGE but can perform editing utilizing
expression basis and physical buffers of fitted morphable DECA model.

Methods	3D RGB Radiance	3D Semantic Radiance	GAC	LAC	SGE	FID ↓ (256^2)	KID $(\times 10^3)$ ↓ (256^2)	#Pars
GRAM	✓	✗	✗	✗	✗	17.36	30.82	1.95 M
FE-NerF	✓	✓	✓	✗	✓	28.97	39.33	31.09 M
SemFaceEdit (Ours)	✓	✓	✓	✓	✓	19.81	32.81	8.78 M
SemFaceEdit (Ours) (Single Stage)	✓	✓	✓	✓	✓	21.27	34.73	8.78 M
SofGAN*	✗	✓	✓	✓	✓	25.72	30.86	30.68 M
Diffusion-Rig*	✗	✗	✓	✗	✗	16.84	30.64	144.6 M
Co-Diff†	✗	✗	✓	✗	✓	14.63	27.43	459.6 M

Diff-Rig first estimates facial priors by fitting Deca model [10] on the given image, then employs a diffusion-based approach to generate view-consistent facial images in 2D space. Changing these buffers along with diffusion noise allows for appearance manipulation and editing, but this methodology lacks semantic-guided editing by design. Sof-GAN produces high-resolution, multi-view facial appearances using a 2D GAN approach, but its training relies on semantically labeled 3D scans. However, the use of 2D GANs prevents it from achieving 3D interpretability. The method presented in GRAM introduces an effective point sampling technique on learnable 2D manifolds. Nonetheless, it falls short in estimating semantic radiance and achieving disentanglement between geometry and appearance. FE-NerF predicts the semantic field but falls short in terms of disentanglement and control over semantic appearance.

Our 3D inversion outcomes, showcased in Figure 7(b,e), are compared against FE-NerF depicted in Figure 7(a,d), respectively. Notably, the semantic inversion by FE-NerF struggles with intricate semantics and produces noisy results, as demonstrated in Figure 7(a). In contrast, our method achieves successful 3D inversion for such cases. Our approach further enables semantic appearance control on the generated RGB-radiance fields. After obtaining optimized geometry latent z and appearance latents z_i through inversion process, we can directly manipulate appearances by changing latent codes z_i corresponding to a semantic region i, as demonstrated in Figure 7(g). In contrast to our approach which relies on pivotal inversion, Diff-Rig generate the images, shown in Figure 7(c,f) by first estimating facial priors by fitting Deca face model [10] on the input image. Note that for Diff-Rig, we only present the output generated after Stage-1 in Figure 7. More accurate results can be obtained by fine-tuning the model on a personalized album, however, this requires multi-view and diverse images of the same face.
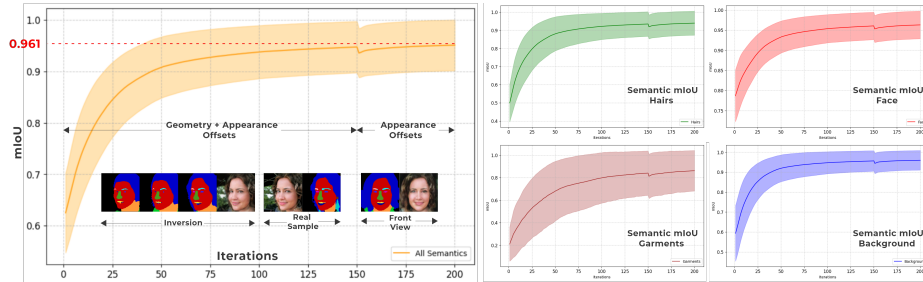
Fig. 9. During the optimization process for 3D inversion on 1000 CelebAMask-HQ Dataset images, we track the Mean Intersection over Union (mIoU). Embedded images qualitatively illustrate the inversion progression of an image in a challenging pose. Furthermore, we track the Mean Intersection over Union (mIoU) of all $n = 4$ semantics, namely: Hairs, Face, Garments, and Background.

Quantitative comparison. Table 1 presents the assessment of image quality using the Frechet Inception Distance (FID) [13] and Kernel Inception Distance (KID) [4]. This evaluation is performed between 10K generated images from randomly sampled latent codes and an equal number of randomly selected real images. We significantly improve the neural rgb-radiance generation compared to that of FE-NerF, as evident in Table 1 while jointly predicting the semantic radiance field. For Diffusion-Rig, we use 10K real images from the CelebAMask-HQ dataset. We estimate buffers (pose, normal, and albedo) by fitting the Deca Model [10] and generate outputs from stage-1, as proposed by [9]. In contrast, for Co-Diff, we input 10K semantic masks from the CelebAMask-HQ dataset for facial image generation to calculate metrics. To assess the 3D inversion capability of SemFaceEdit, we perform inversions on 1K images from the CelebAMask-HQ Dataset, calculating the mean Intersection over Union (mIoU) for all n semantics. Across all inverted images, our approach converges to 0.85 mIoU within 50 iterations. The per-iteration mIoU and semantic class-wise mIoU progressions are depicted in Figure 9.

Ablation Studies. We conducted several ablation studies to determine the optimal choice for SemFaceEdit architecture. In Table 1, we show improved generated image metrics after second-stage training. This process involves refining Appearance Module weights while keeping Geometry module weights frozen. The *Suppl. Material* contains extra ablation studies examining (1) the impact of shared Appearance Modules and (2) Depth Ablation.

Limitations. Our appearance module uses the generated semantic radiance from the geometry module through a semantic volume masking layer. The generation of RGB-radiance is constrained to adhere to the densities of the semantic radiance, which limits the creation of detailed and fine-grained geometry. Additionally, any discrepancies in the semantic geometry propagates to final facial appearance. We include illustrations of such cases in *supp. material*. The high computational requirements for training generative NeRFs limit the proposed

method's ability to generate high-quality images, resulting in image quality that lags behind traditional methods using 2D GANs and DDPMs. In the future, more efficient point sampling and implicit representations could be explored to reduce computational overhead and achieve higher quality radiances.

5   Conclusion

This work introduces a novel approach for semantic face editing on generative radiance manifolds. We enable effective editing capabilities by simultaneously estimating each point's semantic information and RGB radiance. The key to our approach lies in utilizing a differentiable Semantic Volume Masking layer, which effectively partitions points according to their predicted semantics. This partitioning allows us to condition the RGB radiance with latent codes, resulting in a finely tuned level of control over the editing process. With SemFaceEdit, we successfully attain comprehensive control over geometry and appearance, including transferring distinct facial attributes between images. Through experiments, we demonstrate the capability of our method to effectively control the visual attributes of distinct semantics while preserving other regions by manipulating latent codes. Moreover, SemFaceEdit achieves enhanced GAN inversion and faster convergence compared to existing state-of-the-art methods.

References

1. Abdal, R., Zhu, P., Mitra, N.J., Wonka, P.: Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. ACM Transactions on Graphics (ToG) **40**(3), 1–21 (2021)
2. An, S., Xu, H., Shi, Y., Song, G., Ogras, U., Luo, L.: Panohead: Geometry-aware 3d full-head synthesis in 360. arXiv preprint arXiv:2303.13071 (2023)
3. Athar, S., Shu, Z., Samaras, D.: Flame-in-nerf: Neural control of radiance fields for free view face animation. In: 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG). pp. 1–8. IEEE (2023)
4. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. arXiv preprint arXiv:1801.01401 (2018)
5. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L.J., Tremblay, J., Khamis, S., et al.: Efficient geometry-aware 3d generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16123–16133 (2022)
6. Chan, E.R., Monteiro, M., Kellnhofer, P., Wu, J., Wetzstein, G.: 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. pp. 5799–5809 (2021)
7. Chen, A., Liu, R., Xie, L., Chen, Z., Su, H., Yu, J.: Sofgan: A portrait image generator with dynamic styling. ACM Transactions on Graphics **41**(1) (2022)
8. Deng, Y., Yang, J., Xiang, J., Tong, X.: Gram: Generative radiance manifolds for 3d-aware image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10673–10683 (2022)

9. Ding, Z., Zhang, X., Xia, Z., Jebe, L., Tu, Z., Zhang, X.: Diffusionrig: Learning personalized priors for facial appearance editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12736–12746 (2023)

10. Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3d face model from in-the-wild images. ACM Transactions on Graphics (ToG) **40**(4), 1–13 (2021)

11. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)

12. Gu, J., Liu, L., Wang, P., Theobalt, C.: Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. arXiv preprint arXiv:2110.08985 (2021)

13. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)

14. Huang, Z., Chan, K.C., Jiang, Y., Liu, Z.: Collaborative diffusion for multi-modal face generation and editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6080–6090 (2023)

15. Jiang, K., Chen, S.Y., Liu, F.L., Fu, H., Gao, L.: Nerffaceediting: Disentangled face editing in neural radiance fields. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–9 (2022)

16. Jo, K., Shim, G., Jung, S., Yang, S., Choo, J.: Cg-nerf: Conditional generative neural radiance fields. arXiv preprint arXiv:2112.03517 (2021)

17. Kim, G., Kwon, T., Ye, J.C.: Diffusionclip: Text-guided diffusion models for robust image manipulation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2426–2435 (2022)

18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

19. Lee, C.H., Liu, Z., Wu, L., Luo, P.: Maskgan: Towards diverse and interactive facial image manipulation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)

20. Leimkühler, T., Drettakis, G.: Freestylegan: Free-view editable portrait rendering with the camera manifold. arXiv preprint arXiv:2109.09378 (2021)

21. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4d scans. ACM Trans. Graph. **36**(6), 194–1 (2017)

22. Liao, Y., Schwarz, K., Mescheder, L., Geiger, A.: Towards unsupervised learning of generative models for 3d controllable image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2020)

23. Mescheder, L., Geiger, A., Nowozin, S.: Which training methods for gans do actually converge? In: International conference on machine learning. pp. 3481–3490. PMLR (2018)

24. Michalkiewicz, M., Pontes, J.K., Jack, D., Baktashmotlagh, M., Eriksson, A.: Implicit surface representations as layers in neural networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4743–4752 (2019)

25. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)

26. Niemeyer, M., Geiger, A.: Giraffe: Representing scenes as compositional generative neural feature fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11453–11464 (2021)

27. Oechsle, M., Peng, S., Geiger, A.: Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5589–5599 (2021)
28. Or-El, R., Luo, X., Shan, M., Shechtman, E., Park, J.J., Kemelmacher-Shlizerman, I.: Stylesdf: High-resolution 3d-consistent image and geometry generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13503–13513 (2022)
29. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2337–2346 (2019)
30. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3d face model for pose and illumination invariant face recognition. In: 2009 sixth IEEE international conference on advanced video and signal based surveillance. pp. 296–301. Ieee (2009)
31. Roich, D., Mokady, R., Bermano, A.H., Cohen-Or, D.: Pivotal tuning for latent-based editing of real images. ACM Transactions on graphics **42**(1), 1–13 (2022)
32. Sitzmann, V., Thies, J., Heide, F., Nießner, M., Wetzstein, G., Zollhofer, M.: Deepvoxels: Learning persistent 3d feature embeddings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2437–2446 (2019)
33. Sun, J., Deng, Q., Li, Q., Sun, M., Ren, M., Sun, Z.: Anyface: Free-style text-to-face synthesis and manipulation. In: Proceedings of the ieee/cvf conference on computer vision and pattern recognition. pp. 18687–18696 (2022)
34. Sun, J., Wang, X., Shi, Y., Wang, L., Wang, J., Liu, Y.: Ide-3d: Interactive disentangled editing for high-resolution 3d-aware portrait synthesis. ACM Transactions on Graphics (ToG) **41**(6), 1–10 (2022)
35. Sun, J., Wang, X., Zhang, Y., Li, X., Zhang, Q., Liu, Y., Wang, J.: Fenerf: Face editing in neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7672–7682 (2022)
36. Tucker, R., Snavely, N.: Single-view view synthesis with multiplane images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 551–560 (2020)
37. Yu, A., Li, R., Tancik, M., Li, H., Ng, R., Kanazawa, A.: Plenoctrees for real-time rendering of neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5752–5761 (2021)
38. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Bisenet: Bilateral segmentation network for real-time semantic segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 325–341 (2018)
39. Zheng, Y., Abrevaya, V.F., Bühler, M.C., Chen, X., Black, M.J., Hilliges, O.: Im avatar: Implicit morphable head avatars from videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13545–13555 (2022)
40. Zhou, P., Xie, L., Ni, B., Tian, Q.: Cips-3d: A 3d-aware generator of gans based on conditionally-independent pixel synthesis. arXiv preprint arXiv:2110.09788 (2021)
41. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. arXiv preprint arXiv:1805.09817 (2018)
42. Zhu, P., Abdal, R., Qin, Y., Wonka, P.: Sean: Image synthesis with semantic region-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5104–5113 (2020)
43. Zhuang, Y., Zhu, H., Sun, X., Cao, X.: Mofanerf: Morphable facial neural radiance field. In: European Conference on Computer Vision. pp. 268–285. Springer (2022)