Supplementary material for 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

1 Loss Functions

As discussed in main paper, we train our network in two stages in an adversarial learning setting, using non-saturating GAN loss with R1 regularization [8]. 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, using the loss function given in Equation 4. Our discriminators have similar architectures as utilized by [2], with minor changes. To ensure that the final images generated by the Appearance Module Ψ_a are in the correct pose, we introduce an additional branch consisting of linear layer within discriminator D_c to estimate the pose θ. We also add an additional branch consisting of linear layer within D_s to estimate latent code that generated the semantic radiance. The adversarial losses for discriminating semantics and images is defined in Equation 1

$$\begin{aligned}\mathcal{L}(D_s, D_c, \Psi_a, \Psi_g) = &\mathrm{E}_{z\sim p_z^g, \theta\sim p_\theta}[f(D_s(\Psi_g(z,\theta))] \\ &+ \mathrm{E}_{z_i\sim p_{z_i}^a, \theta\sim p_\theta}[f(D_s(\Psi_a(z_i,\theta))] \\ &+ \mathrm{E}_{I\sim p_{rim}}[f(-D_c(I)) + \lambda_{im}||\nabla D_c(I)||^2] \\ &+ \mathrm{E}_{L\sim p_{rsem}}[f(-D_s(L)) + \lambda_s||\nabla D_s(L)||^2] \end{aligned} \tag{1}$$

Here, $f(x) = log(1 + exp(x))$ is softplus function, I is a real image sample from p_{rim} and corresponding annotated label $L \sim p_{rsem}$. Furthermore, to tackle concave geometry challenges and improve GAN inversion, we integrate two additional losses, namely, pose loss and latent loss outlined in Equations 2 and 3, respectively. These are combined with the loss presented in Equation 1. Additionally, it's worth noting that Ψ_g learns to generate L with N labels, while Ψ_a conditions clubbed semantics defined by n.

$$\begin{aligned}\mathcal{L}_p = &\mathrm{E}_{z_i\sim p_{z_i}^a, \theta\sim p_\theta}||f(D_c^p(\Psi_a(z_i,\theta)) - \theta||^2 \\ &+ \mathrm{E}_{I\sim p_{rim}}||f(D_c^p(I) - \hat{\theta}||^2 \end{aligned} \tag{2}$$

$$\mathcal{L}_{latent} = \mathrm{E}_{z\sim p_z^g, \theta\sim p_\theta} ||f(D_s^l(\Psi_g(z,\theta)) - z||^2 \tag{3}$$

During first stage of training we incorporate all of the defined losses as given by Equation 4.

$$\mathcal{L}_{S1} = \mathcal{L}(D_s, D_c, \Psi_a, \Psi_g) + \lambda_p \mathcal{L}_p + \lambda_l \mathcal{L}_{latent} \tag{4}$$

Proceeding to second stage of training, we freeze weights of Geometry Module Ψ_g and fine-tune the weights that affect appearance. The effective loss function in second stage is outlined in equation 5.

$$\mathcal{L}_{S2} = \mathcal{L}(D_c, \Psi_a) + \lambda_p \mathcal{L}_p \tag{5}$$

where $\mathcal{L}(D_c, \Psi_a)$ is given by:

$$\begin{aligned}\mathcal{L}(D_c, \Psi_a) = \mathrm{E}_{z_i\sim p_{z_i}^a, \theta\sim p_\theta}[f(D_s(\Psi_a(z_i,\theta))] \\ + \mathrm{E}_{I\sim p_{rim}}[f(-D_c(I)) + \lambda_{im}||\nabla D_c(I)||^2]\end{aligned} \tag{6}$$

In our experiments, we empirically set the values of $\lambda_{im} = 5, \lambda_s = 1, \lambda_p = 10$ and $\lambda_l = 1$, during first stage training. During second stage we use $\lambda_{im} = 1$ and $\lambda_p = 10$.

2 Pivotal Inversion

In this section we present mathematical formulation for optimizing δw^+ and δw_i^+ to obtain geometric and appearance latent codes $w \in \mathcal{W}$ and $w_i \in \mathcal{W}_i$. To edit an input image obtained either from $\mathrm{I} \sim p_{rim}$ (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 [9]. In the main paper, we described that 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. Similarly, for appearance, we optimize for δw_i^+ to make the appearance resemble the given image I.

To achieve the inversion of a real image I, we estimate the geometric and appearance offsets, δw^+ and δw_i^+ respectively, as articulated in Equation 7.

$$\begin{aligned}\mathcal{L}(\delta w^+, \delta w_i^+) = \lambda_s \mathcal{L}_s(S', S) \\ + \lambda_{im}\mathcal{L}_{im}(I', I) + \lambda_{vgg}\mathcal{L}_{vgg}(I', I)\end{aligned} \tag{7}$$

Here, $\mathcal{L}_s$ can be Cross-Entropy Loss or Mean squared Error Loss between predicted and ground truth semantics. $\mathcal{L}_{im}$ is Mean squared error loss and $\mathcal{L}_{vgg}$ is perceptual loss [6] with VGG [10] as backbone network for feature extraction. In our experiments, we setup $\mathcal{L}_s$ as Mean square error loss and $\lambda_s = 10, \lambda_{im} = 1$, and $\lambda_{vgg} = 1$.

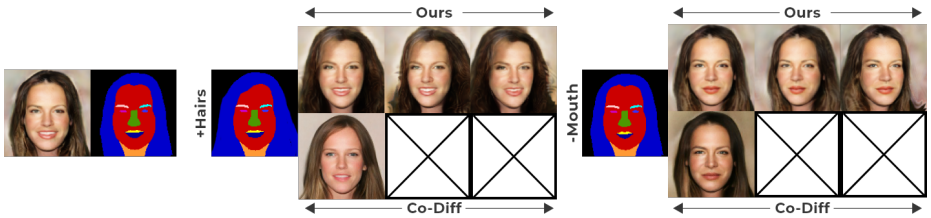

Fig. 1. Effect of manipulation in the semantic mask by expanding the hair region semantics and the effect of shrinking at the mouth region compared with Co-Diff [5].

2.1 Editing Faces and Facial Attribute Transfer

When editing faces using a semantic mask or transferring semantic attributes represented by a mask $\mathcal{M}_k$, as detailed in the main paper, let r denote the region requiring alteration while preserving the integrity of the remaining image portion. For editing generated image I' given an edited semantic mask S_{ed} corresponding to original semantics S. We modify inversion equation as presented in Equation 7, following the formulation specified in Equation 8, to estimate δw^+ and δw_i^+ for the required edit. It's important to note that we only apply loss within the known regions, i.e., $(1-r)$, while granting the generator the flexibility to realistically fill the region r, thereby achieving the intended editing outcome. In the context of transferring facial attributes between a given source image I_s and a target image I_t, the inversion equation, articulated in Equation 7, is similarly adjusted in line with the formulation presented in Equation 8. This adjustment serves the purpose of estimating δw^+ and δw_i^+, necessary for the attribute transfer. In our experiments, we setup $\mathcal{L}_s$ as Mean square error loss and $\lambda_s = 10, \lambda_{im} = 1$, and $\lambda_{vgg} = 1$.

$$\begin{aligned}\mathcal{L}(\delta w^+, \delta w_i^+) = \lambda_s \mathcal{L}_s(S', S_{ed}) \\ + \lambda_{im}\mathcal{L}_{im}(I' \odot (1-r), I \odot (1-r)) \\ + \lambda_{vgg}\mathcal{L}_{vgg}(I' \odot (1-r), I \odot (1-r))\end{aligned} \tag{8}$$

$$\begin{aligned}\mathcal{L}(\delta w^+, \delta w_i^+) = \lambda_s \mathcal{L}_s(S' \odot (1-r), S_s \odot (1-r)) \\ + \lambda_s \mathcal{L}_s(S' \odot r, S_t \odot r) \\ + \lambda_{im}\mathcal{L}_{im}(I' \odot (1-r), I_s \odot (1-r)) \\ + \lambda_{vgg}\mathcal{L}_{vgg}(I' \odot (1-r), I_s \odot (1-r)) \\ + \lambda_{im}\mathcal{L}_{im}(I' \odot r, I_t \odot r) \\ + \lambda_{vgg}\mathcal{L}_{vgg}(I' \odot r, I_t \odot r)\end{aligned} \tag{9}$$

3 Qualitative Results

In Figure 3, we show neural 72 rendering of generated semantic and rgb-radiances produced by latent codes randomly sampled from a gaussian distribution. Fur-

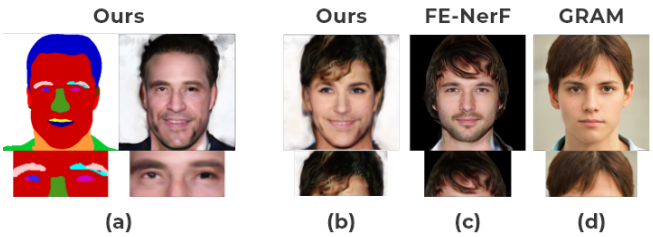

Fig. 2. Artifacts due to discrepancy in semantic radiance (a) and Hair quality obtained by different methods (c), (d) compared with Ours (b).

thermore, we provide a link to supplementary video containing GIFs showcasing multi-view renderings from different camera viewpoints: `https://youtube.com/shorts/b53EkPVK328?feature=share`

3.1 Semantics Guided Editing

We present editing obtained by our method compared with Co-Diff [5] in Figure 1. Note that since our approach generates semantic and rgb-radiances in a volume, by design it supports generation of multi-view images by changing camera position. In contrast Co-Diff [5] employs a diffusion-based model to generate image in 2D space.

3.2 Limitations

The generation of RGB-radiance is influenced by the points grouped by the semantic volume masking layer. Consequently, any discrepancies in the generated semantic radiance propagate, leading to visual artifacts in the final image, as shown in Figure 2(a). Additionally, the generation of RGB-radiance is constrained to adhere to the densities of the semantic radiance, which limits the creation of fine-grained geometry, such as hairs. We observe that the quality of hair generation in methods that learn the RGB-radiance field in conjunction with semantic radiances lags behind other 3D-aware GAN methods that do not incorporate semantics [3]. For example, in Figure 2(b) and Figure 2(c), both our method and FE-NerF [11], respectively, produce hair strands that are less realistic compared to those generated by Gram [3], as illustrated in Figure 2(c).

4 Data Preprocessing

We utilize the CelebAMask-HQ dataset [7] for training the proposed network because it includes semantic segmentation masks along with facial images. Initially, we crop and align all images from the dataset to center the facial region

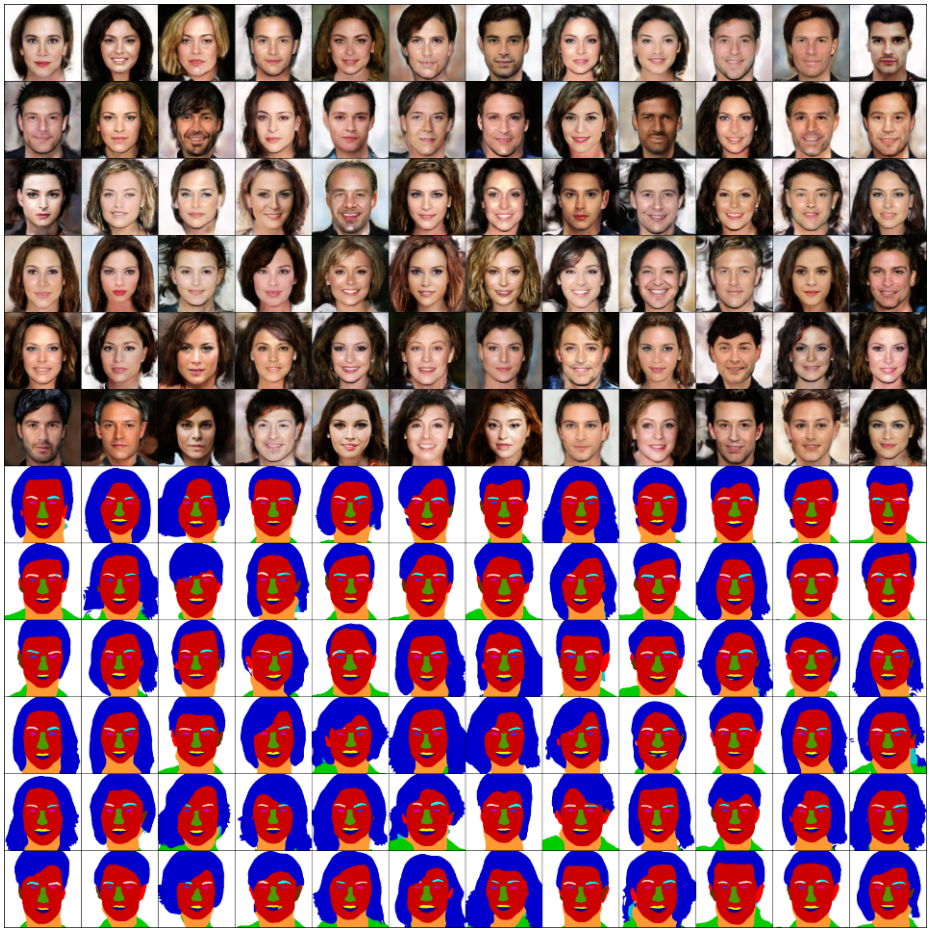

Fig. 3. Renderings of Semantic-radiance and RGB-radiance on image space generated by our approach by random latent code $\mathbf{z} \in \mathrm{R}^d$ and $\mathbf{z}_i \in \mathrm{R}^d$.

using the method described in [1]. We also adjust the semantic masks according to the alignment transformation of the corresponding images. Furthermore, for each face, we estimate pose similar to [4]. Alignment ensures that facial features (e.g., eyes, nose, mouth) are consistently positioned across all training images. This consistency along with camera pose information helps the model learn more accurate and generalized implicit representations of facial features and their spatial relationships in volumetric field.

Training on in the wild images. To enhance diversity in terms of ethnicity and appearance, the model should be trained on a more varied dataset. When using images sourced from the web or repositories, the facial images must first be aligned as previously discussed. Subsequently, a semantic face parsing network,

Table 1. Comparing FID and KID between 5K generated and 5K real CelebA-MaskHQ [7] Dataset images with different weight sharing in Appearance Module. Each variant is trained for 60K iterations (2 epoch) jointly learning weights for both Geometry Module and Appearance Module (First stage).

Method	**FID** ↓	**KID** ($\times 10^3$) ↓
No shared weights	30.36	41.62
Fully Shared AM (shared Linear Color layers)	24.43	36.32
Proposed Sharing (SemFaceEdit)	22.65	34.74

Table 2. Comparing FID and KID between 5K generated and 5K real CelebA-MaskHQ [7] Dataset images with different architecture depth of Geometry Module and Appearance Module. Each variant is trained for 60K iterations (2 epochs), jointly learning weights for both Geometry Module and Appearance Module (First stage).

Depth			
Geometric Module	**Appearance Module**	**FID** ↓	**KID** ($\times 10^3$) ↓
4	4	27.18	38.35
6	6	24.12	35.43
8	8	22.65	34.74

such as those described in [7][13][12] can be employed to create the semantic masks necessary for training.

5 Ablation Studies

In the proposed architecture *(Figure* 3 *of main paper)*, we implement shared weights for the Appearance Module (excluding Linear Color Layers) across all semantic categories. This design choice not only reduces model complexity but also contributes to enhanced metrics, as demonstrated in Table 1. Further, we perform experiments by varying the number of FiLM Layers [2] in both the Geometry and Appearance Module. It's noteworthy that we observe improvements in performance corresponding to an increase in the number of layers within the neural networks, as illustrated in Table 2. In consideration of computational constraints and model complexity, the results presented from SemFaceEdit are based on a depth value of 8.

References

1. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In: Proceedings of the IEEE international conference on computer vision. pp. 1021–1030 (2017)
2. 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)
3. 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)
4. Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., Tong, X.: Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In: IEEE Computer Vision and Pattern Recognition Workshops (2019)
5. 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)
6. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. pp. 694–711. Springer (2016)
7. 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)
8. 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)
9. 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)
10. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
11. 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)
12. Yu, C., Gao, C., Wang, J., Yu, G., Shen, C., Sang, N.: Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. International journal of computer vision **129**, 3051–3068 (2021)
13. Zheng, Y., Yang, H., Zhang, T., Bao, J., Chen, D., Huang, Y., Yuan, L., Chen, D., Zeng, M., Wen, F.: General facial representation learning in a visual-linguistic manner. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18697–18709 (2022)