My3DGen: A Scalable Personalized 3D Generative Model

Luchao Qi[1], Jiaye Wu[2], Annie N. Wang[1], Shengze Wang[1], and Roni Sengupta[1]

[1] University of North Carolina at Chapel Hill
{lqi, awang13, shengzew, ronisen}@cs.unc.edu
[2] University of Maryland, College Park
jiayewu@cs.umd.edu

Abstract. In recent years, generative 3D face models (e.g., EG3D) have been developed to tackle the problem of synthesizing photo-realistic faces. However, these models are often unable to capture facial features unique to each individual, highlighting the importance of personalization. Some prior works have shown promise in personalizing generative face models, but these studies primarily focus on 2D settings. Also, these methods require both fine-tuning and storing a large number of parameters for each user, posing a hindrance to scalability. Another challenge of personalization is the limited number of training images available for each individual, which often leads to overfitting when using full fine-tuning methods. Our proposed approach, My3DGen, generates a personalized 3D prior of an individual using as few as 50 training images. My3DGen allows for novel view synthesis, semantic editing of a given face (e.g. adding a smile), and synthesizing novel appearances, all while preserving the original person's identity. We decouple the 3D facial features into global features and personalized features by freezing the pre-trained EG3D and training additional personalized weights through low-rank decomposition. As a result, My3DGen introduces only **240K** personalized parameters per individual, leading to a **127**× reduction in trainable parameters compared to the **30.6M** required for fine-tuning the entire parameter space. Despite this significant reduction in storage, our model preserves identity features without compromising the quality of downstream applications.

Keywords: Personalization · 3D-GAN · 3D Face

1 Introduction

Recently, dramatic advancements in deep generative models like generative adversarial networks (GANs) [5,34,56,59] and diffusion models [20,30] have led to a surge in their popularity for computer vision applications. Notably, GANs have proven to be particularly powerful at generating realistic photos of faces [35–39]. These techniques have been extended to 3D vision as well, leading to the development of models that are capable of reconstructing 3D models of existing 2D facial images, synthesizing novel appearances, and editing various facial attributes such as facial expressions [12,50,51,55,65]. These models play a crucial role in enhancing the authenticity of virtual communication, AR/VR/MR, and content creation, thereby increasing engagement.

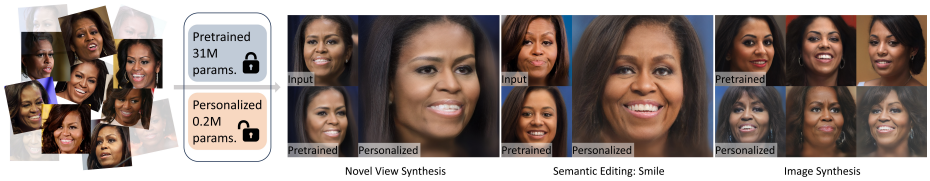

Fig. 1: Given 50 images of *Michelle Obama*, we personalize a pre-trained 3D generative prior and demonstrate the applications in various downstream tasks. Each downstream task presents the original input image of *Michelle* (top left), alongside the corresponding output generated using the pre-trained face prior (bottom left), compared to the output using our personalized face prior (right). Our personalized prior can faithfully retain the key facial characteristics of *Michelle Obama*, as opposed to the pre-trained prior.

However, current 3D generative models [12, 50, 51, 55, 65] are unable to create authentic 3D face models of particular subjects. Although these models can synthesize photo-realistic fake faces, they cannot generate, reconstruct, or modify the distinctive traits of a particular real person's face without distorting their identity. This problem is further exaggerated for underrepresented demographics whose facial characteristics are sparsely represented in widely used training datasets like FFHQ [38] or CelebA [47]. Existing inversion techniques [1, 6, 40, 61, 70, 74] are only able to preserve identity by tuning the model separately for every test image, an impractical and inefficient approach. Furthermore, they cannot edit the inverted image or synthesize novel appearances without identity distortion. Thus, this paper presents an approach to create a personalized 3D generative prior for an individual. This prior enables 3D facial reconstruction, synthesis of novel appearances, and editing of existing appearances, while maintaining the individual's identity.

A major obstacle to the deployment of personalized 3D generative models at scale for real-world applications is their enormous storage demand. We can illustrate this scalability problem with a concrete example: Consider that naively personalizing a 3D generative model such as EG3D [12] requires storing approximately 31 million parameters (121 MB) for each user. For three billion users (monthly active users of Facebook), this would require 363 PB of memory, an extremely cost-prohibitive demand even for a company the size of Meta. Evidently, we need to design more parameter-efficient personalization techniques that will enable the building of 3D generative priors for a large population. Another challenge of personalization is that an average user often takes only a limited number of photos of themselves each day or week [49]. The limited training set size makes it difficult to fine-tune a large global generative model, often leading to overfitting, mode collapse [2], and data drift [44, 48], all of which prevent the model from generalizing to unseen test images of an individual.

Our idea is to decouple the facial features of an individual into (a) shared global features that can be represented by a generative model trained across many different identities and (b) personalized features of an individual that can be represented with much fewer trainable parameters, trained on images of that individual only. We use EG3D [12] as our pre-trained generative model and train additional weights for low-rank decompositions of every convolutional

and fully-connected layer to capture the personalized features of that individual. Drawing inspiration from the recent success of Low-Rank Adaption (LoRA) [32] in parameter-efficient fine-tuning of large language models [33, 77] and diffusion models [64], we use LoRA during personalization, which hasn't been well explored in convolution-heavy GAN-based models before. Our approach allows for personalization using only 240K (0.9 MB) trainable parameters instead of requiring fine-tuning of all 31 million (121 MB) parameters. For three billion users, this means that only 2.7 PB of storage memory would be needed instead of 363 PB.

Quantitative and qualitative analyses show that our proposed My3DGen outperforms the pre-trained 3D generative model EG3D [12] on multiple tasks including 3D reconstruction, novel appearance synthesis, image enhancement, and semantic editing. Additionally, our personalized model can produce results similar to those achieved by naively fine-tuning a pre-trained model with 31 million parameters [12], while only using as few as 240K trainable parameters. We further provide insights based on different parametrization strategies. We observe that increasing the rank of LoRA modules only contributes to better overfitting of the background or small stylistic changes without improving the shape or the identity-preserving performance after personalization. Furthermore, we find that personalizing the earlier layers of StyleGAN2 has the most impact on the resulting quality, indicating that coarse and middle layers are more responsible for capturing the shape and identity of the person.

In conclusion, we propose the following contributions. (1) To the best of our knowledge, our research presents the first attempt to personalize a 3D generative model and demonstrates its effectiveness in various downstream tasks; (2) We design a generative prior where an individual's facial features are disentangled into global features–represented by a pre-trained 3D generative model, and personalized features–captured by training additional low-rank weights requiring only 240K stored parameters. This method helps to avoid overfitting and has the potential to improve performance over naive tuning without LoRA; (3) Departing from previous works that primarily utilized LoRA solely for linear layers within transformers or diffusion models, our approach incorporates LoRA into convolution-centric GAN architectures, presenting an innovative perspective on using LoRA for 3D personalization.

2 Related Work

3D Face Reconstruction. Parametric 3D models, also known as 3D Morphable Models (3DMM) [8, 24], are often widely used for modeling 3D human faces through a linear combination of basic shapes. Typically, 3DMMs are constructed using high-quality facial scans from multiple individuals. However, when fitting a 3DMM to a test image [42, 68, 69], the result often appears unrealistic due to limitations in the expressiveness of the linear blended model. Also, training a 3DMM using 2D photo collections is challenging due to the absence of corresponding 3D scans. As a result, recent research has focused on developing generative models for 3D faces [13, 27, 55], which are able to achieve a remarkable level of facial detail and capture small wrinkles and bumps by training solely on

2D images. Nevertheless, these models only aim to generate arbitrary fake faces, and there is a lack of research addressing the personalization of such models, which is the focus of our paper.

Personalized Generative Models. Personalization plays a significant role in the field of generative AI, with diverse applications including deepfakes, video avatars, talking heads, and text-to-image (T2I) generation [2, 15, 16, 57, 63, 64, 73, 76]. Previous inversion works [3, 6, 61] designed a model-dependent encoder or fine-tuned the pre-trained model to best fit an input image. However, these methods do not possess the ability to generate diverse appearances or edit existing appearances of an individual without distorting their identity. Most inversion methods also require updating and storing a separate set of network weights for every single image, which is impractical. In personalizing a 3D generative model, our goal is to use only a single set of network weights that can reconstruct, edit, and synthesize any novel appearances of an individual while preserving identity. Our work is inspired by the recent success in 2D generative model personalization [52, 79]. However, 2D personalization cannot create 3D faces or produce unique perspectives, and frequently falters for non-frontal positions due to a lack of geometric information. Our paper extends personalization from 2D to 3D by assuming that an individual's facial appearance can be decomposed into global features represented by the pre-trained model and personalized characteristic features represented by parameter-efficient low-rank adaptive weights.

Parameter-Efficient Fine-Tuning. Large foundation models [10, 22, 46, 60, 62] often achieve impressive performance for tasks in their domain. However, the huge number of parameters in such models often prevents them from being fine-tuned for downstream tasks using a limited budget. For example, GPT-4 [54] contains 1.76 trillion parameters, which is impractical for most users to fine-tune. To tackle this issue, many parameter-efficient fine-tuning (PEFT) techniques have been previously proposed to fine-tune models efficiently. In natural language processing, approaches [31, 32, 45] have been proposed to enable efficient adaptation of pre-trained language models to various downstream applications without fine-tuning all of the model's parameters. For image generation, ControlNet [80], HyperDreamBooth [64], and AnimateDiff [28] have been proposed to tune pretrained diffusion models with additional modules. Our approach is inspired by LoRA [32], a technique for efficiently finetuning large language foundation models by imposing low-rank structures on weight matrices. LoRA [32] was first proposed as a way to predict additional network weights without changing the pre-trained transformer, allowing efficient adaptation and storage for task-specific models [32]. Subsequently, LoRA has found widespread use in fine-tuning pre-trained networks for various downstream tasks, including network quantization, parameter budget allocation, and continual learning [19, 67, 81].

3 Method

Our objective is to personalize a 3D generative model with ideally 50 images of an individual. First, we introduce background concepts in Sec. 3.1. Then we formulate the problem of personalizing EG3D [12], a pre-trained 3D generative model, in Sec. 3.2. Finally, we discuss how we incorporate parameter-efficient personalization into our model in Section 3.3 with a visual overview in Fig. 2.

3.1 Preliminaries

EG3D. In this work we use the state-of-the-art 3D generative model EG3D [12] as the pre-trained model that captures global facial features across multiple identities. EG3D has four main components: (i) a StyleGAN2 generator that takes as input a random latent code and outputs $256 \times 256 \times 96$ feature maps, which are further reshaped into three $256 \times 256 \times 32$ triplanes, (ii) a neural renderer that decodes triplane features and renders a face given a camera pose input, (iii) a super-resolution module that upsamples rendered images to a resolution of 512×512, and (iv) a StyleGAN2 discriminator that differentiates between generated images and real images. The entire pipeline is trained following the typical non-saturating minimax GAN loss [26]. In this paper, we discard the discriminator and will use a reconstruction loss to personalize the generator.

LoRA. Our approach is based on Low-Rank Adaptation of Large Language Models (LoRA) [32], a technique for efficiently finetuning convolution-free transformer networks by imposing low-rank structures on weight matrices. LoRA shows that a pre-trained model's weight matrix $W_0 \in \mathbb{R}^{d\times k}$ for any fully-connected layer can be fine-tuned with a low-rank decomposition using the following formulation: Let $W_{\text{ft}} = W_0 + \Delta W = W_0 + BA$ be the fine-tuned adapted weight matrix, where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ are trained while W_0 is frozen, and the rank $r \ll \min(d, k)$ [32]. Although this approach demonstrates impressive decomposition performance for transformers [29] and diffusion models [64], previous work solely focuses on the decomposition of linear layers [28, 41, 64, 67, 72], while GAN-based models are convolution-heavy. In our paper, we apply LoRA to the convolution-heavy EG3D model for personalization.

3.2 Personalization Formulation

We consider a scenario where an individual has a training set of N 2D images x_i, each with an associated camera pose c_i, denoted as $\mathcal{D}_p = \{(x_i, c_i)\}_{i=1}^{N}$. We let $G(\cdot; \theta_G + \Delta\theta_G)$ denote our personalized EG3D model where G is the EG3D architecture, θ_G represents the frozen pretrained weights, and $\Delta\theta_G$ represents our trainable parameters. Our goal is to tune this model on a low-dimensional manifold in the $\mathbf{W}$ latent space, dubbed a personal convex hull [52], which is a subspace defined by all latent codes of the N images. Our personalization scheme is inspired by the success of PTI [61] and MyStyle [52] in tuning a pretrained 2D StyleGAN. We first invert images from $(x_i, c_i) \sim \mathcal{D}_p$ with associated camera poses into latent vectors w_i, dubbed anchors in MyStyle [52], with an off-the-shelf inversion technique [39]. That is, we freeze both the camera pose and the model weights while only optimizing the randomly initialized latent vector.

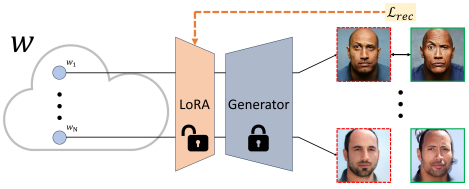

Fig. 2: Architecture of our personalization approach. We project an individual's images into StyleGAN2's $\mathcal{W}$ space through latent code optimization to obtain a set of latent anchors. We then tune the generator to reconstruct an individual's images. During tuning, the generator is frozen while only LoRA weights are personalized.

We then tune the weights $\Delta\theta_G$ of the model G using reconstruction objective $\mathcal{L}_{rec}$, which represents the difference between each image x_i and its reconstruction. Formally,

$$\begin{aligned}\mathcal{L}_{rec}^{(i)} = &\mathcal{L}_{lpips}\left(G\left(w_i, c_i; \theta_G + \Delta\theta_G\right), x_i\right) \\ &+ \lambda_{\mathcal{L}_2} \left\| G\left(w_i, c_i; \theta_G + \Delta\theta_G\right) - x_i \right\|_2 \end{aligned} \tag{1}$$

and we have $\Delta\theta_{G_p} = \text{argmin}_{\Delta\theta_G} \sum_{i=1}^{N} \mathcal{L}_{rec}^{(i)}$ as our optimized parameters for the personalized EG3D model G_p.

3.3 Parameter Efficient Personalization

Naively fine-tuning the model can lead to forgetting of knowledge learned in pretraining and entanglement of pre-trained facial features with the personalized features, compromising both the model's interpretability and generalization. Therefore, in order to decouple the global features, captured in the pre-trained model, from the personalized features, we freeze the weights of the pre-trained model and tune additional weights to capture an individual's distinct facial priors. We use the technique of Low-Rank Adaptation (LoRA) [32] to train the additional personalized weights, using only 240K parameters per identity.

The original LoRA paper focuses on convolution-free transformers where linear fully connected layers are decomposed into low-rank submatrices A and B, respectively [32]. However, EG3D's generator StyleGAN2 and super-resolution modules are convolution-heavy, and convolution operations are often implemented with general matrix multiplication using the well-known 'im2col' algorithm [4]. Thus, when adapting for personalization, we decompose both the convolution and fully connected layers in the StyleGAN2 generator and super-resolution modules using LoRA [3]. Pretrained weights W_0 of the StyleGAN2 generator and the super-resolution module are frozen while only A and B are trained. We tune all the parameters of the neural renderer in EG3D, as it is relatively small with 4K parameters. Under this setting, the number of parameters needed to be tuned is determined by the rank r. With rank $r = 1$, we can reduce the 30.6M trainable parameters of EG3D to only 240k, a reduction of 127×.

4 Experiments

We first discuss the details of our evaluation framework including datasets, experiment pipeline, and evaluation metrics in Sec. 4.1. Then we present both qualitative and quantitative results of our approach, My3DGen, in Sec. 4.2. We further analyze the effect of LoRA for personalization in Sec. 4.3. Finally, we present ablation studies in Sec. 4.4.

4.1 Experiment Details

Dataset. We conduct experiments using facial images of celebrities, the same dataset used by Mystyle [52]. Images are preprocessed following [12,35] to align and crop faces to 512 × 512. Finally, the faces are separated into reference sets and test sets for each celebrity. The number of images included in the reference set and the test set for each celebrity is presented in the supplementary. Unless

[3] We go into detail about our implementation of the convolutional LoRA decomposition in the supplementary material.

otherwise noted, for each celebrity, we personalize a model using 50 images from the reference set as the training set.

We use an off-the-shelf pose extraction [18] model to both identify the face region and label the pose of the face, following the same pipeline in EG3D [12]. We project images into the model's $\mathcal{W}$ latent space following the StyleGAN [39] optimization scheme for 500 iterations to obtain their latent codes (anchors).

Training. We choose 50 images per individual for all personalization tunings. The effect of dataset size is further discussed in Sec. 4.4. We tune the model starting from the pretrained EG3D (on FFHQ), which outputs images at resolutions of both 128×128 ($\boldsymbol{I}_{128}$) and 512×512 ($\boldsymbol{I}_{512}$). Hyper-parameters are chosen as follows: $\lambda_{lpips} = \lambda_{\mathcal{L}_2} = 1$. We apply $\mathcal{L}_{rec}$ on both $\boldsymbol{I}_{128}$ and $\boldsymbol{I}_{512}$. Our further experiments show that a higher LPIPS regularization weight ($\lambda_{lpips} > 1$) will lead to checkerboard-style artifacts and a lower LPIPS weight ($\lambda_{lpips} < 0.1$) will cause nonphotorealistic artifacts, which aligns with results in previous work [6]. Based on this observation, we use $\lambda_{lpips} = 1$.

Evaluations. We evaluate our methods for the following primary downstream tasks: 1) *Image inversion* through PTI [61]; 2) *Image interpolation* where we interpolate between two anchors in the latent space and generate images that morph from one face to another; 3) *Image synthesis* where the goal is to generate novel appearances of an individual by sampling from a latent space, following the protocol outlined in Mystyle [52]; 4) *Image enhancement* tasks such as image-inpainting and super-resolution; 5) *Semantic editing* where the goal is to modify the facial expression or age of the person while maintaining the identity and pose. Note that we modify the model weights after personalization via PTI only for image-inversion tasks, for consistency with prior research on 3D GAN [12]. For any novel views, we render faces with a yaw range of ±0.35 (radians) and a pitch range of ±0.25 (radians) relative to the frontal face for all our experiments. Unless otherwise specified, we evaluate the performance of the personalized model on unseen test images. As our experiments were conducted in a 3D environment, we strongly suggest the reader refer to the supplementary video materials.

Metrics. When evaluating inversion outcomes, it is standard to use pixel-based metrics such as PSNR and SSIM or neural network-based perceptual metrics such as LPIPS [82] and DISTS [21] to assess the inverted image and compare it with the original image. However, due to errors in the estimation of face poses, Live3Dportrait [70] reports that a minor misalignment between the ground truth and the estimated face poses will cause traditional pixel-based metrics to be unreliable. Recent work further indicates that deep perceptual image metrics are also sensitive to small misalignments [25]. Therefore, we additionally adopt the facial identity score ID_{sim}, *i.e.* a metric to evaluate the preservation of the individual's identity [11, 52]. Nonetheless, we still include DISTS and LPIPS results in our inversion tasks to align with prior works [12, 70].

The identity score ID_{sim} of an image is determined by the individual's reference set, which contains all the images available from that individual. Given an image, we extract the identity features and report the cosine similarity of the

given image to the nearest image in the personal reference set. Formally,

$$\mathrm{ID}_{sim}\left(w_i, c_i\right) = max\left\{\langle R(G(w_i, c_i, \cdot)), R(x_j)\rangle\right\}_{j=1}^{N} \tag{2}$$

where R is a pretrained ArcFace [17] network for feature recognition, $\langle\cdot,\cdot\rangle$ computes the cosine similarity between its argument as the ID scores [58], and N is the number of 2D images in a reference set. We report ID_{sim} for personal identity preservation evaluation in all of our downstream applications, specifically in synthesis and interpolation tasks where there is no ground truth.

4.2 Applications of My3DGen

Inversion. Given a 2D RGB image of a face, image inversion here refers to performing 3D reconstruction of the face. We apply PTI [61] for inversion to align with the original EG3D work [12]. We optimize a randomly initialized latent code for 600 iterations, followed by fine-tuning the model for an additional 350 iterations [70]. We calculate DISTS and LPIPS for single-view inversion, and report ID_{sim} to evaluate multi-view reconstruction. As shown in Fig. 3, the pretrained EG3D can have artifacts such as i) a visible seam between the face and the rest of the head (1st row, 3rd column), ii) 3D shape distortion (4th row, 5th column), and iii) identity drift caused by changing pose, which can be observed in Fig. 1 regarding novel view synthesis. Furthermore, the pretrained EG3D overly smooths skin textures. We recommend readers to zoom in to observe these finer details (2nd column of Fig. 3). In contrast, ours decomposes facial features into global features and personalized features, producing a higher quality and identity-preserving inversion. Compared to full fine-tuning, our approach yields almost the same quality for single-view inversion as measured by LPIPS, and produces similar multiview inversion results with a slightly higher ID_{sim}. We speculate that full fine-tuning can result in better single-view quality by

Fig. 3: Qualitative evaluation for image inversion, *i.e.* generating 3D-aware view synthesis from a single input image. Ours is My3DGen with LoRA rank $r = 1$ where the number of trainable parameters = 0.2M. Visual differences are highlighted with a red-box, zoom in to view finer details.

Table 1: Quantitative evaluation for image inversion. F.T. indicates fine-tuning, and Personal. Params. indicates the number of personalized parameters. DISTS and LPIPS compare the inverted image against the input image in the same pose, ID_{sim} evaluates the preservation of identity across multiple poses.

F.T.	LoRA	DISTS ↓	LPIPS ↓	$\mathrm{ID}_{\mathrm{sim}}$ ↑	Personal. Params. ↓
-	-	0.12	0.19	0.56	-
✓	-	0.08	0.12	0.60	31M
✓	✓	0.08	0.13	0.61	0.2M

inverting images with an overfitted background, but may also introduce artifacts by overfitting particular facial features. For background overfitting, we examine this problem in Fig. 8 where we demonstrate that increasing LoRA rank enhances background fitting but provides minor improvements in terms of ID_{sim}. For facial features, as shown in Fig. 3 (5th row), full fine-tuning does not preserve the identity in the lower jaw and ear regions, which are obscured in the input image. Personalizing the model with LoRA can help reduce these artifacts, even including rendering 'floaters' from NeRF [75]. We suggest the readers check the inversion video results in the supplementary for more examples.

Interpolation. The results obtained from the inversion tasks are overfitted to the test image via PTI, requiring separate optimization of the generator for each image. Our approach aims to provide a single generative model for all images of an individual without requiring further optimization. We support this claim by generating images from latent vectors produced through linear interpolation between two randomly selected training anchors [52]. We interpolate between the anchor pair at 10 equally spaced interpolation weights. At each interpolation step θ, given the interpolated latent code, we randomly generate 20 novel views and compute the average ID_{sim} scores of each view as described in Eq. 2. We repeat this process for 10 randomly selected anchor pairs for each personalized model, averaging our results across all models and pairs. We compare the results before and after personalization in Fig. 4 and show that personalization maintains the identity when traversing between two anchor images (exemplified by *Scarlett Johansson*) while the pretrained model distorts identity.

Synthesis. We conduct image synthesis to further examine the personalization capacity of our approach. Our goal is to produce new, distinctive images of an individual that have not been seen before. We sample a latent code from the convex hull follow-

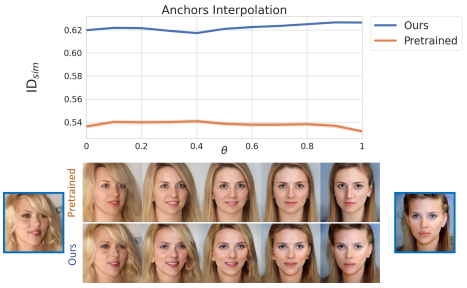

Fig. 4: Quantitative (top) and qualitative (bottom) evaluation for interpolation in latent space between two anchor images, highlighted in color. We measure identity preservation using ID_{sim}, for which illustrative visual results at different interpolation steps θ are also provided. We compare the results before and after personalization.

Fig. 5: Qualitative evaluation for synthesizing novel appearances of an individual. Generated images of *Oprah Winfrey* are provided for visual inspection.

Table 2: Quantitative evaluation for novel appearance synthesis. F.T. indicates fine-tuning and Personal. Params. indicates the number of personalized parameters.

F.T.	LoRA	ID_{sim} ↑	Diversity ↑	Personal. Params. ↓
-	-	0.53	0.19	-
✓	-	0.62	0.21	31M
✓	✓	0.62	0.21	0.2M

ing MyStyle [52]. Then we feed the latent code into the generator together with a random pose. To this end, we randomly synthesize images from each generator and compare the results based on identity preservation. We evaluate the image synthesis results and present them in Table 2, accompanied by visual examples in Fig. 5. To assess the diversity of the synthesized images, we follow the protocol proposed by Ojha et al. [53]. Specifically, for each celebrity, we produce 1,000 images and assign each of them to one of the k training images, by choosing the one with the lowest LPIPS distance. We then compute the standard deviation of pairwise LPIPS distances within members of the same cluster and then average over the k clusters. Our model surpasses pretrained EG3D in all metrics and performs comparably to full fine-tuning with fewer trainable parameters.

Image Enhancement. We choose the tasks of image inpainting (IP) and super-resolution (SR) as representative examples of image enhancement. To perform inpainting, we utilize a mask positioned at the center of the faces and feed the masked images into the generator. We subsequently post-process the resulting outputs by blending the generator's output within the masked area with the input in other areas. The background regions of the outputs are replaced through post-processing using a Lanczos-upsampled version of the input image. These regions have been segmented according to the methodology in [71]. For SR, we reduce the original image size from 512×512 to 32×32 and supply the blurred images to the generator. It is noteworthy that in previous studies on image enhancement [83], no quality evaluation is performed against a reference standard, *i.e.* a ground truth image. "Quality" here refers to a pixel-to-pixel comparison against the original ground truth. The reasoning for this is that although the restored facial details may differ from those of the original image, personalization still restores the key facial characteristics of the individual, resulting in valid restorations [52, 79]. We adhere to the established convention and do not include quality evalu-

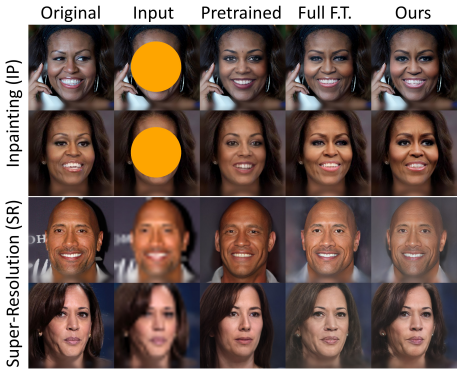

Fig. 6: Qualitative evaluation for image enhancement by inpainting (IP) and super-resolution (SR). The original images are degraded as input images and then fed into the model. We have included original images for the benefit of readers who may not be familiar with the faces of *Michelle Obama, Dwayne Johnson, and Kamala Harris.*

Table 3: Quantitative evaluation for image enhancement. F.T. indicates fine-tuning and Personal. Params. indicates the number of personalized parameters. User % reflects the percentages of responses for each option in the image enhancement tasks.

	F.T.	LoRA	ID_{sim} ↑	User % ↑	Personal. Params. ↓
IP	-	-	0.62	12.7	-
	✓	-	0.72	-	31M
	✓	✓	0.72	81.8	0.2M
SR	-	-	0.61	7.9	-
	✓	-	0.73	-	31M
	✓	✓	0.73	88.5	0.2M

ations either. Therefore for completeness, in addition to using the ID_{sim} metric, we further perform user studies that rely on subjective assessments of identity preservation in the experiments. We used Amazon Mechanical Turk to gather 330 responses from 17 users. In the study, users were shown the original image, along with two results: one from the pretrained model and the other from the personalized model, in a randomized order. Users were instructed to select the result that most closely resembled the person in the picture and maintained high fidelity to the original input. In cases where both results were similar, participants could select 'No Preference'. Results are reported as a percentage of each selected option. Images utilized in the user study were selected from a random subset of those used for quantitative evaluation. We present the enhancement results in Fig. 6 and Table 3, showing visual examples of inpainting (IP) and super-resolution (SR) tasks followed by both quantitative and qualitative analysis. The User % for 'No Preference' is 5.5% for inpainting and 3.6% for super-resolution. Our results demonstrate that My3DGen successfully integrates personal features into the pretrained EG3D with both higher quantitative score (ID_{sim}), and qualitative preference (User %).

Semantic Editing. Besides enhancing identity preservation in reconstruction tasks, personalization also enables tailored editing of facial attributes. For example, while a generalized pre-trained model may learn how an average human smiles or ages with time, it is not able to capture the specific smile or aging process of an individual well. Our hypothesis is that a personalized generative prior can enable us to more accurately depict how specific attributes of an individual change. To validate this hypothesis, we perform semantic editing of 'smile' and 'age' features using 3D generative priors, both pretrained and personalized. We start by identifying suitable editing directions for each model using the InterFaceGAN framework [66], followed by reconstructing the input image and then conducting personalized semantic editing in the α-space, as described in MyStyle [52]. Our findings are illustrated in Table 4, where we also employ the user study as quantitative metrics for evaluation. Our study design for the user survey aligns with

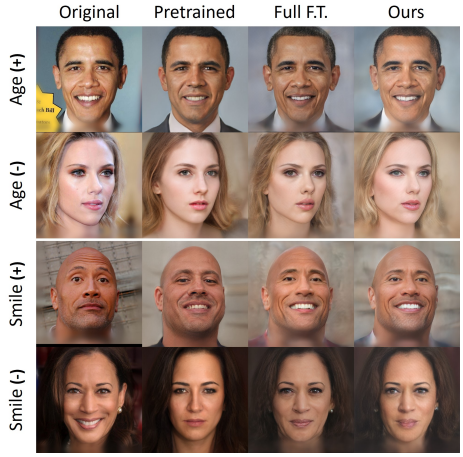

Fig. 7: Qualitative evaluation for semantic editing of 'smile' and 'age'. We show *Barack Obama, Scarlett Johansson, Dwayne Johnson, and Kamala Harris* top-to-bottom.

Table 4: Quantitative evaluation for semantic editing. F.T. indicates fine-tuning and Personal. Params. indicates the number of personalized parameters. User % reflects the percentages of responses for each option.

F.T.	LoRA	ID_{sim} ↑	User % ↑	Personal. Params. ↓
-	-	0.60	4.8	-
✓	-	0.76	-	31M
✓	✓	0.76	95.2	0.2M

that of the image enhancement evaluations, including 165 responses from a total of 17 participants. Quantitatively, our model yields a nearly identical ID_{sim} to the fully fine-tuned EG3D model and substantially higher ID_{sim} than the pre-trained model. Qualitatively, we note that our model can sometimes outperform the fully fine-tuned model in retaining image styles, such as lighting, background color, and hairstyle post-editing. A detailed examination of Fig. 7 demonstrates improved preservation of skin tone and background color in the first row, enhanced reproduction of hair color, style, and background color in the second row, and better handling of lighting and skin tone in the last row, showing LoRA may mitigate overfitting in semantic editing.

4.3 Analysis of Personalization with LoRA

In Sec. 4.2, we show that personalizing 3D-GAN with LoRA can help avoid overfitting and has the potential to improve downstream results such as inversion and semantic editing compared to naive full fine-tuning. In this section, we start by demonstrating that tuning the pre-trained model with LoRA rank $r = 1$ suffices for personalization. We then study the importance of feature blocks with different resolutions for personalization with rank 1.

Effect of Rank in LoRA. To investigate the impact of rank selection in LoRA and find the optimal rank, we repeat the same personalization method for each celebrity, with LoRA ranks of 1, 4, and 16, tuning the generator accordingly. We report the model's performance on inversion and interpolation tasks in Table 5. For inversion tasks, increasing the rank from 1 to 16 doesn't help improve identity preservation measured by ID_{sim}, but it does help to fit the background and thus leads to a better LPIPS score, for which visual results are shown in Fig. 8.

Table 5: Quantitative analysis of the rank of LoRA for inversion and interpolation tasks. '-' indicates full fine-tuning. We report the average ID_{sim} across latent paths from the interpolation task.

	Inversion		Interpolation	Personal.
LoRA	LPIPS ↓	ID_{sim} ↑	ID_{sim} ↑	Params. ↓
-	0.12	0.60	0.62	31M
rank=1	0.13	0.60	0.62	0.2M
rank=4	0.11	0.60	0.63	0.9M
rank=16	0.09	0.60	0.63	3.5M

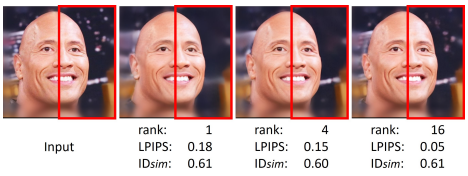

Fig. 8: Visual examples of inversion results for different LoRA ranks. The backgrounds are highlighted to aid inspection along with LPIPS and ID_{sim} scores.

This clarifies the contrast between our method and full fine-tuning regarding single-view inversion results outlined in the inversion task in Sec. 4.2. However, matching backgrounds is not the primary focus of personalization, so the slight performance tradeoff is acceptable in this case. Additionally, there is only a slight improvement in ID_{sim} for interpolation tasks when the rank increases. Thus, we choose to use rank = 1 in most of our experiments.

Personalization of Feature Blocks. We show that feature blocks with different resolutions have different levels of importance during personalization. Determining the importance of weights in a network is an open question [7]. Previous work proposes including gradient information to evaluate the importance of LoRA modules during the tuning [81]. Here, after tuning with LoRA, we follow the common approach in model pruning where the change in parameter magnitude is used as a criterion to evaluate layer importance [23, 43]. That is, $|\Delta W|/|W_0| \times 100\%$, where ΔW is the personalized LoRA weights and W_0 is the pre-trained weights. In EG3D, the first 7 resolution blocks of StyleGAN2 are used as the backbone to generate tri-plane representations of 256×256 resolution. For each resolution block, we calculate the relative LoRA weight change percentage compared to the pre-trained weight.

We report the mean and variance of the changes across different individuals in Fig. 9. The results show that feature blocks of resolution 8×8, 16×16, and 32×32 require more weight changes than other resolution blocks. This indicates that personal facial features are learned through the 'coarse' and 'middle' layers during personalization while little changes are required for 'fine' layers. This finding aligns with the results in StyleGAN2 [38] that 'fine' layers affect only micro-features and finer details in the image such as color schemes and hairstyles.

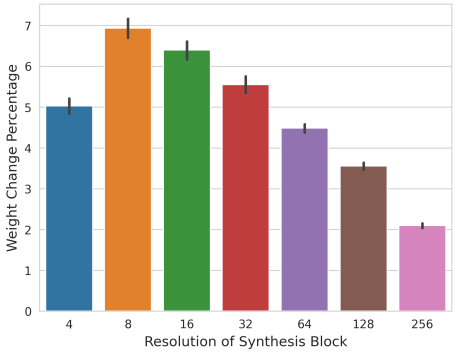

Fig. 9: LoRA weight change compared to the pre-trained weights after personalization, averaged across celebrities.

4.4 Ablation study

Effect of Dataset Size. We investigate the effect of training dataset size on the performance of the personalized generator. We sample various random subsets of images for each celebrity, with sizes of 10, 50, and 100, and tune the generator accordingly. Next, we assess the generator's performance using the interpolation tasks outlined in Section 4.2, which evaluates the generalization capacity of the personal convex hull by traversing through the latent subspace. We report the average ID_{sim} across latent paths.

Table 6: The effect of training dataset size $\mathcal{D}_p$ on personalization. We evaluate personalization through the interpolation task in Sec. 4.2 and report the average ID_{sim}.

Dataset $\mathcal{D}_p$ Size Ablations	Metric	$\mathcal{D}_p = 10$	$\mathcal{D}_p = 50$	$\mathcal{D}_p = 100$
	$\text{ID}_{\text{sim}} \uparrow$	0.618	0.628	0.629

As shown in Table 6, the performance improves significantly from 10 to 50 images, but there is no significant improvement from 50 to 100 images. While further experiments are required, we speculate that adding more images may not contribute to dataset diversity and might even hurt the results [52]. Therefore, we use 50 images for our personalization method in most of our experiments, unless otherwise stated.

AFHQv2 Cats. In addition to human faces, we also extend our personalization method to cat faces. Leveraging a photo album consisting of 22 in-the-wild images of one individual cat, we detect poses following [9] and apply the same procedure used for human faces to personalize the pretrained EG3D-AFHQ model, which was pretrained on a dataset of 15000 animal images, including 5000 cat images across different identities and breeds.

The results, showcased in Fig. 10, demonstrate that our personalization technique significantly enhances the quality of the pretrained triplane representations for cat faces. This successful extension of our approach demonstrates the versatility and effectiveness of our method across different domains, paving the way for personalized 3D generative modeling of full human bodies, other animals, or objects.

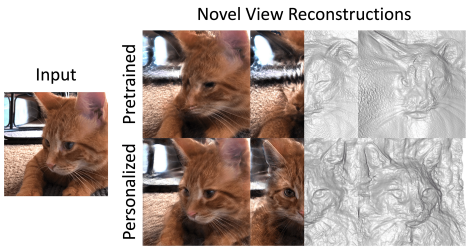

Fig. 10: Comparison between a pre-trained model and a personalized model for inverting an in-the-wild cat photo.

5 Discussion

Limitations and Future Work. Our model can accurately capture facial features but encounters difficulty when objects heavily obscure the face (e.g. hats, phones, etc). Further works may leverage the power of EG3D-based encoder [6, 78] for a better inversion performance. Ours also struggles with heavily cropped faces, where the whole face is not fully captured in the original image and thus image boundaries are filled with reflection-padded values to align the face during preprocessing. One may mask out these invalid values during inversion. This will be similar to the in-painting task discussed in Section 4.2.

Ethical Considerations. This study holds the potential to generate manipulated images of actual individuals, posing a substantial societal threat. Future research for detecting fake composites is needed.

Conclusion. We propose a parameter-efficient framework to personalize a large pretrained 3D generative model. Ours incorporates an individual's personal facial features into the pretrained model for downstream applications while preserving unique identity features. This will enable scalable personalization of 3D generative models in the real world.

References

1. Abdal, R., Lee, H.Y., Zhu, P., Chai, M., Siarohin, A., Wonka, P., Tulyakov, S.: 3davatargan: Bridging domains for personalized editable avatars. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4552–4562 (2023) 2
2. Aghabozorgi, M., Peng, S., Li, K.: Adaptive IMLE for Few-shot Pretraining-free Generative Modelling (2023) 2, 4
3. Alaluf, Y., Patashnik, O., Cohen-Or, D.: Restyle: A residual-based stylegan encoder via iterative refinement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6711–6720 (2021) 4
4. Anderson, A., Vasudevan, A., Keane, C., Gregg, D.: High-Performance Low-Memory Lowering: GEMM-based Algorithms for DNN Convolution. In: 2020 IEEE 32nd International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD). pp. 99–106 (Sep 2020). https://doi.org/10.1109/SBAC-PAD49847.2020.00024, https://ieeexplore.ieee.org/abstract/document/9235051?casa_token=fNkRHtQ8Gj0AAAAA:XStPIeWHOx-UfmbHwnurKBI8e9HXGTfD5M-zUNQJydPncbdRj2k5AMEepQAmaevg2piyfsd70A, iSSN: 2643-3001 6
5. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 214–223. PMLR (06–11 Aug 2017), https://proceedings.mlr.press/v70/arjovsky17a.html 1
6. Bhattarai, A.R., Nießner, M., Sevastopolsky, A.: Triplanenet: An encoder for eg3d inversion. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2024) 2, 4, 7, 14
7. Blalock, D., Ortiz, J.J.G., Frankle, J., Guttag, J.: What is the State of Neural Network Pruning? (Mar 2020). https://doi.org/10.48550/arXiv.2003.03033, http://arxiv.org/abs/2003.03033, arXiv:2003.03033 [cs, stat] 13
8. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques. p. 187–194. SIGGRAPH '99, ACM Press/Addison-Wesley Publishing Co., USA (1999). https://doi.org/10.1145/311535.311556, https://doi.org/10.1145/311535.311556 3
9. Brad: kairess/cat_hipsterizer (Mar 2024), https://github.com/kairess/cat_hipsterizer, original-date: 2018-10-12T15:00:15Z 14
10. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 1877–1901. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf 4
11. Cao, K., Rong, Y., Li, C., Tang, X., Loy, C.C.: Pose-Robust Face Recognition via Deep Residual Equivariant Mapping. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5187–5196. IEEE, Salt Lake City, UT, USA (Jun 2018). https://doi.org/10.1109/CVPR.2018.00544, https://ieeexplore.ieee.org/document/8578642/ 7

12. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L., Tremblay, J., Khamis, S., Karras, T., Wetzstein, G.: Efficient Geometry-aware 3D Generative Adversarial Networks. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16102–16112. IEEE, New Orleans, LA, USA (Jun 2022). https://doi.org/10.1109/CVPR52688.2022.01565, https://ieeexplore.ieee.org/document/9880428/ 1, 2, 3, 4, 5, 6, 7, 8
13. Chan, E.R., Monteiro, M., Kellnhofer, P., Wu, J., Wetzstein, G.: pi-GAN: Periodic Implicit Generative Adversarial Networks for 3D-Aware Image Synthesis. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5795–5805. IEEE, Nashville, TN, USA (Jun 2021). https://doi.org/10.1109/CVPR46437.2021.00574, https://ieeexplore.ieee.org/document/9577547/ 3
14. Chellapilla, K., Puri, S., Simard, P.: High performance convolutional neural networks for document processing. In: Tenth international workshop on frontiers in handwriting recognition. Suvisoft (2006) 1
15. Cheng, K., Cun, X., Zhang, Y., Xia, M., Yin, F., Zhu, M., Wang, X., Wang, J., Wang, N.: Videoretalking: Audio-based lip synchronization for talking head video editing in the wild. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–9 (2022) 4
16. Choi, J.M., Christman, M., Sengupta, R.: Personalized Video Relighting With an At-Home Light Stage (Dec 2023). https://doi.org/10.48550/arXiv.2311.08843, http://arxiv.org/abs/2311.08843, arXiv:2311.08843 [cs] 4
17. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019) 8
18. 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: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 0–0 (2019) 7
19. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: QLoRA: Efficient Finetuning of Quantized LLMs (May 2023), http://arxiv.org/abs/2305.14314, arXiv:2305.14314 [cs] 4
20. Dhariwal, P., Nichol, A.: Diffusion Models Beat GANs on Image Synthesis (Jun 2021), http://arxiv.org/abs/2105.05233, arXiv:2105.05233 [cs, stat] 1
21. Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image Quality Assessment: Unifying Structure and Texture Similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(5), 2567–2581 (May 2022). https://doi.org/10.1109/TPAMI.2020.3045810, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence 7
22. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (Jun 2021), http://arxiv.org/abs/2010.11929, arXiv:2010.11929 [cs] 4
23. Elesedy, B., Kanade, V., Teh, Y.W.: Lottery Tickets in Linear Models: An Analysis of Iterative Magnitude Pruning (Jul 2021), http://arxiv.org/abs/2007.08243, arXiv:2007.08243 [cs, stat] 13
24. Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3d face model from in-the-wild images. ACM Trans. Graph. **40**(4) (jul 2021). https://doi.org/10.1145/3450626.3459936, https://doi.org/10.1145/3450626.3459936 3

25. Ghildyal, A., Liu, F.: Shift-tolerant perceptual similarity metric. In: European Conference on Computer Vision. pp. 91–107. Springer (2022) 7
26. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020) 5
27. Gu, J., Liu, L., Wang, P., Theobalt, C.: StyleNeRF: A Style-based 3D-Aware Generator for High-resolution Image Synthesis (Oct 2021), http://arxiv.org/abs/2110.08985, arXiv:2110.08985 [cs, stat] 3
28. Guo, Y., Yang, C., Rao, A., Wang, Y., Qiao, Y., Lin, D., Dai, B.: AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning (Jul 2023), http://arxiv.org/abs/2307.04725, arXiv:2307.04725 [cs] 4, 5
29. He, X., Li, C., Zhang, P., Yang, J., Wang, X.E.: Parameter-efficient model adaptation for vision transformers. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 817–825 (2023) 5
30. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models (Dec 2020), http://arxiv.org/abs/2006.11239, arXiv:2006.11239 [cs, stat] 1
31. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: International Conference on Machine Learning. pp. 2790–2799. PMLR (2019) 4
32. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-Rank Adaptation of Large Language Models (Oct 2021), http://arxiv.org/abs/2106.09685, arXiv:2106.09685 [cs] version: 2 3, 4, 5, 6, 1
33. Hu, Z., Wang, L., Lan, Y., Xu, W., Lim, E.P., Bing, L., Xu, X., Poria, S., Lee, R.K.W.: LLM-Adapters: An Adapter Family for Parameter-Efficient Fine-Tuning of Large Language Models (Oct 2023), http://arxiv.org/abs/2304.01933, arXiv:2304.01933 [cs] 3
34. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: International Conference on Learning Representations (2018), https://openreview.net/forum?id=Hk99zCeAb 1
35. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive Growing of GANs for Improved Quality, Stability, and Variation (Feb 2018). https://doi.org/10.48550/arXiv.1710.10196, http://arxiv.org/abs/1710.10196, arXiv:1710.10196 [cs, stat] 1, 6
36. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. Advances in neural information processing systems **33**, 12104–12114 (2020) 1
37. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-free generative adversarial networks. Advances in Neural Information Processing Systems **34**, 852–863 (2021) 1
38. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019) 1, 2, 13
39. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8110–8119 (2020) 1, 5, 7
40. Ko, J., Cho, K., Choi, D., Ryoo, K., Kim, S.: 3d gan inversion with pose optimization. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2967–2976 (2023) 2

41. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1931–1941 (2023) 5
42. Lattas, A., Moschoglou, S., Ploumpis, S., Gecer, B., Ghosh, A., Zafeiriou, S.: AvatarMe++: Facial Shape and BRDF Inference With Photorealistic Rendering-Aware GANs. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(12), 9269–9284 (Dec 2022). https://doi.org/10.1109/TPAMI.2021.3125598, https://www.computer.org/csdl/journal/tp/2022/12/09606538/1ymEN8wBXRC, publisher: IEEE Computer Society 3
43. Lee, J., Park, S., Mo, S., Ahn, S., Shin, J.: Layer-adaptive sparsity for the Magnitude-based Pruning (May 2021), http://arxiv.org/abs/2010.07611, arXiv:2010.07611 [cs] 13
44. Lee, J., Cho, K., Kiela, D.: Countering Language Drift via Visual Grounding (Sep 2019). https://doi.org/10.48550/arXiv.1909.04499, http://arxiv.org/abs/1909.04499, arXiv:1909.04499 [cs] 2
45. Lin, Z., Madotto, A., Fung, P.: Exploring versatile generative language model via parameter-efficient transfer learning. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 441–459. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.findings-emnlp.41, https://aclanthology.org/2020.findings-emnlp.41 4
46. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach (2019) 4
47. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision. pp. 3730–3738 (2015) 2
48. Lu, Y., Singhal, S., Strub, F., Courville, A., Pietquin, O.: Countering Language Drift with Seeded Iterated Learning. In: Proceedings of the 37th International Conference on Machine Learning. pp. 6437–6447. PMLR (Nov 2020), https://proceedings.mlr.press/v119/lu20c.html, iSSN: 2640-3498 2
49. Miller, A.D., Edwards, W.K.: Give and take: a study of consumer photo-sharing culture and practice. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 347–356. ACM, San Jose California USA (Apr 2007). https://doi.org/10.1145/1240624.1240682, https://dl.acm.org/doi/10.1145/1240624.1240682 2
50. Nguyen-Phuoc, T., Li, C., Theis, L., Richardt, C., Yang, Y.L.: Hologan: Unsupervised learning of 3d representations from natural images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7588–7597 (2019) 1, 2
51. Niemeyer, M., Geiger, A.: GIRAFFE: Representing Scenes as Compositional Generative Neural Feature Fields. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11448–11459. IEEE, Nashville, TN, USA (Jun 2021). https://doi.org/10.1109/CVPR46437.2021.01129, https://ieeexplore.ieee.org/document/9577414/ 1, 2
52. Nitzan, Y., Aberman, K., He, Q., Liba, O., Yarom, M., Gandelsman, Y., Mosseri, I., Pritch, Y., Cohen-Or, D.: Mystyle: A personalized generative prior. ACM Transactions on Graphics (TOG) **41**(6), 1–10 (2022) 4, 5, 6, 7, 9, 10, 11, 14, 1, 2, 3
53. Ojha, U., Li, Y., Lu, J., Efros, A.A., Lee, Y.J., Shechtman, E., Zhang, R.: Few-shot image generation via cross-domain correspondence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10743–10752 (2021) 10

54. OpenAI: GPT-4 Technical Report (Mar 2023). https://doi.org/10.48550/arXiv.2303.08774, http://arxiv.org/abs/2303.08774, arXiv:2303.08774 [cs] 4
55. 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) 1, 2, 3
56. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019) 1
57. Paruchuri, A., Liu, X., Pan, Y., Patel, S., McDuff, D., Sengupta, S.: Motion Matters: Neural Motion Transfer for Better Camera Physiological Measurement (Nov 2023), http://arxiv.org/abs/2303.12059, arXiv:2303.12059 [cs] 4
58. Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D.: Styleclip: Text-driven manipulation of stylegan imagery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2085–2094 (2021) 8
59. Radford, A., Metz, L., Chintala, S.: Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks (Jan 2016), http://arxiv.org/abs/1511.06434, arXiv:1511.06434 [cs] 1
60. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019) 4
61. Roich, D., Mokady, R., Bermano, A.H., Cohen-Or, D.: Pivotal tuning for latent-based editing of real images. ACM Transactions on graphics (TOG) **42**(1), 1–13 (2022) 2, 4, 5, 7, 8, 1
62. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 4
63. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023) 4
64. Ruiz, N., Li, Y., Jampani, V., Wei, W., Hou, T., Pritch, Y., Wadhwa, N., Rubinstein, M., Aberman, K.: HyperDreamBooth: HyperNetworks for Fast Personalization of Text-to-Image Models (Jul 2023), http://arxiv.org/abs/2307.06949, arXiv:2307.06949 [cs] 3, 4, 5
65. Schwarz, K., Liao, Y., Niemeyer, M., Geiger, A.: GRAF: Generative Radiance Fields for 3D-Aware Image Synthesis. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 20154–20166. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper_files/paper/2020/file/e92e1b476bb5262d793fd40931e0ed53-Paper.pdf 1, 2
66. Shen, Y., Gu, J., Tang, X., Zhou, B.: Interpreting the latent space of gans for semantic face editing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9243–9252 (2020) 11
67. Smith, J.S., Hsu, Y.C., Zhang, L., Hua, T., Kira, Z., Shen, Y., Jin, H.: Continual Diffusion: Continual Customization of Text-to-Image Diffusion with C-LoRA (Apr 2023), http://arxiv.org/abs/2304.06027, arXiv:2304.06027 [cs] 4, 5
68. Tran, A.T., Hassner, T., Masi, I., Medioni, G.: Regressing Robust and Discriminative 3D Morphable Models with a Very Deep Neural Network. In: 2017 IEEE

Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1493–1502. IEEE, Honolulu, HI (Jul 2017). https://doi.org/10.1109/CVPR.2017.163, http://ieeexplore.ieee.org/document/8099646/ 3
69. Tran, L., Liu, X.: On learning 3d face morphable model from in-the-wild images. IEEE transactions on pattern analysis and machine intelligence **43**(1), 157–171 (2019) 3
70. Trevithick, A., Chan, M., Stengel, M., Chan, E., Liu, C., Yu, Z., Khamis, S., Chandraker, M., Ramamoorthi, R., Nagano, K.: Real-time radiance fields for single-image portrait view synthesis. ACM Transactions on Graphics (TOG) **42**(4), 1–15 (2023) 2, 7, 8, 3
71. Wadhwa, N., Garg, R., Jacobs, D.E., Feldman, B.E., Kanazawa, N., Carroll, R., Movshovitz-Attias, Y., Barron, J.T., Pritch, Y., Levoy, M.: Synthetic depth-of-field with a single-camera mobile phone. ACM Transactions on Graphics (ToG) **37**(4), 1–13 (2018) 10
72. Wang, H., Xiang, X., Fan, Y., Xue, J.H.: Customizing 360-Degree Panoramas through Text-to-Image Diffusion Models (Nov 2023), http://arxiv.org/abs/2310.18840, arXiv:2310.18840 [cs] 5
73. Westerlund, M.: The emergence of deepfake technology: A review. Technology innovation management review **9**(11) (2019) 4
74. Xie, J., Ouyang, H., Piao, J., Lei, C., Chen, Q.: High-fidelity 3d gan inversion by pseudo-multi-view optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 321–331 (2023) 2
75. Xie, Z., Zhang, J., Li, W., Zhang, F., Zhang, L.: S-NeRF: Neural Radiance Fields for Street Views (Mar 2023). https://doi.org/10.48550/arXiv.2303.00749, http://arxiv.org/abs/2303.00749, arXiv:2303.00749 [cs] 9
76. Yao, S., Zhong, R., Yan, Y., Zhai, G., Yang, X.: DFA-NeRF: Personalized Talking Head Generation via Disentangled Face Attributes Neural Rendering (Jan 2022), http://arxiv.org/abs/2201.00791, arXiv:2201.00791 [cs] 4
77. Yu, D., Naik, S., Backurs, A., Gopi, S., Inan, H.A., Kamath, G., Kulkarni, J., Lee, Y.T., Manoel, A., Wutschitz, L., Yekhanin, S., Zhang, H.: Differentially Private Fine-tuning of Language Models (Jul 2022). https://doi.org/10.48550/arXiv.2110.06500, http://arxiv.org/abs/2110.06500, arXiv:2110.06500 [cs, stat] 3
78. Yuan, Z., Zhu, Y., Li, Y., Liu, H., Yuan, C.: Make Encoder Great Again in 3D GAN Inversion through Geometry and Occlusion-Aware Encoding (Mar 2023), http://arxiv.org/abs/2303.12326, arXiv:2303.12326 [cs] 14
79. Zeng, L., Chen, L., Xu, Y., Kalantari, N.K.: Mystyle++: A controllable personalized generative prior. In: SIGGRAPH Asia 2023 Conference Papers. SA '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3610548.3618171, https://doi.org/10.1145/3610548.3618171 4, 10
80. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3836–3847 (October 2023) 4
81. Zhang, Q., Chen, M., Bukharin, A., He, P., Cheng, Y., Chen, W., Zhao, T.: Adaptive Budget Allocation for Parameter-Efficient Fine-Tuning (Mar 2023). https://doi.org/10.48550/arXiv.2303.10512, http://arxiv.org/abs/2303.10512, arXiv:2303.10512 [cs] 4, 13
82. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 586–595. IEEE,

Salt Lake City, UT (Jun 2018). https://doi.org/10.1109/CVPR.2018.00068, https://ieeexplore.ieee.org/document/8578166/ 7

83. Zhao, S., Cui, J., Sheng, Y., Dong, Y., Liang, X., Chang, E.I., Xu, Y.: Large Scale Image Completion via Co-Modulated Generative Adversarial Networks (Mar 2021), http://arxiv.org/abs/2103.10428, arXiv:2103.10428 [cs] 10

Except for this supplemental PDF, we wrap other visual supplemental materials (images, videos, etc.) into an HTML file that can be viewed in *result.html*. We highly recommend readers to refer to the accompanying videos for a comprehensive examination of the visual outcomes.

A Overview of Appendices

Our appendices contain the following additional details:

- Sec. B describes the details of our convolutional LoRA decomposition, where we show the difference between the original implementation and ours. We compare the results visually in Fig. 11, where facial texture is accompanied by checkerboard artifacts using the original LoRA paper's code implementation [32].
- Sec. C provides details on the celebrity dataset and additional information regarding our ablation studies on the effect of dataset size in Sec. 4.4.
- Sec. D shows image inversion results without PTI [61] in Fig. 12. We reproject latent code into personal convex hull following Mystyle [52] for inversion where model weights remain unchanged.
- Sec. E further discusses the interpolation results shown in Sec. 4.2, particularly why the interpolation curve is different from that in Mystyle [52].
- Sec. F describes the detailed hardware configurations and training time for our experiments.
- In Sec. G, we display failure cases for our experiments in Fig. 13.

B LoRA for Convolutional Layer

LoRA [32] is originally defined for matrix multiplication for fully-connected layers. However, convolution operation with C_1 output channels, C_2 input channels, and kernel size of $k \times k$ is often implemented as matrix multiplication with a matrix W under "im2col" [14] transform on the image X. The matrix W has dimension $W \in \mathbb{R}^{C_1 \times C_2 kk}$.

$$\text{Conv}_\theta(X) = W\,\text{im2col}(X) \tag{3}$$

Fig. 11: Following the same personalization pipeline, we compare reconstructed results using the original LoRA code (middle) and our own LoRA implementation (right). The original algorithm introduces idiosyncratic artifacts, such as diagonal stripe patterns. It is recommended to zoom in for finer details, especially around the cheeks and forehead.

Therefore, we can decompose matrix M for convolution layers similar to LoRA. With a rank r LoRA decomposition, let $B \in \mathbb{R}^{C_1 \times r}$ and $A \in \mathbb{R}^{r \times C_2 kk}$, we have the following equation.

$$W = BA \tag{4}$$

We found official LoRA [32] implementation performs the following decomposition. Matrix W is assumed to have dimension $W \in \mathbb{R}^{C_1 k \times C_2 k}$, while B, A have dimension $B \in \mathbb{R}^{C_1 k \times r}$, $A \in \mathbb{R}^{r \times C_2 k}$.

Ours differs from the original LoRA implementation in two ways:

- LoRA showed weight matrix $W \in \mathbb{R}^{C_1 \times C_2 kk}$ that maps from input space $C_2 kk$ to output space C_1 of a layer can have low rank structure. It is unclear if matrix $W \in \mathbb{R}^{C_1 k \times C_2 k}$ has low rank structure.
- We are surprised to find that the official implementation of LoRA directly interprets the memory content of the matrix $W \in \mathbb{R}^{C_1 k \times C_2 k}$ as $W \in \mathbb{R}^{C_1 \times C_2 kk}$ and perform convolution operation. We suspect this is a bug. Even though the matrix W is trainable, we suspect that such implementation has important consequences on performance, as now we are equivalently trying to find a low rank decomposition for a matrix that has no clear meaning, and might not have a low rank structure.

We compare ours with the official implementation of LoRA in Fig. 11, where the official implementation introduces checkerboard artifacts while our implementation is better at keeping the original image content.

C Dataset Size

Using the same dataset in Mystyle [52], we further process the images following the preprocessing pipeline in EG3D [12]. We show the number of images in the reference and test sets in Table 7. In Sec. 4.4, we conduct ablation studies to investigate the effect of the size of the training set. When tuning on 100 images, if the reference set size is below 100, we use all the images in the reference set as the training set, such as 97 for *Dwayne Johnson* and 92 for *Xi Jinping*. Unless otherwise specified, we tune on 50 images for the majority of our personalization experiments.

Table 7: The sizes of the reference and test sets of our dataset.

Celebrity	Reference set size	Test set size
Barack Obama	192	13
Dwayne Johnson	97	12
Joe Biden	200	13
Kamala Harris	110	7
Michelle Obama	279	9
Oprah Winfrey	135	9
Scarlett Johansson	260	13
Taylor Swift	158	9
Xi Jinping	92	15

D Image Inversion without PTI

In Sec. 4.2, we perform image inversion tasks using PTI [61] to align with previous works [12, 70], where PTI requires changing the model weights for the best inversion quality. Nevertheless, we provide inversion results following Mystyle [52] where the model weights remain unchanged and only the latent code is re-projected into the convex hull. As shown in Fig. 12, although personalization helps maintain identity in the inversion tasks, it still lacks facial details for both full fine-tuning and ours, compared to PTI. Further works may design an encoder for EG3D inversion similar to TriPlaneNet [6].

E ID_{sim} Curve Shape in Interpolation Tasks

Interestingly, unlike the previous findings of Mystyle [52], there is no significant difference in ID_{sim} scores between the interpolated latent codes and the anchors. The interpolation ID_{sim} curve in Mystyle follows a reserved U-shape, while our curve is flatter, as shown in Fig. 4. It is hypothesized that this lack of difference may be due to both our ID_{sim} metric design and EG3D's 3D advantage, where the extreme properties of anchors, such as pose, have a smaller impact on ID_{sim} compared to 2D-GANs.

F Training Time

We perform our personalization experiments on 4 NVIDIA RTX A6000 GPUs. Our total training time is 5 hours with LoRA, compared to 6 hours without LoRA.

G Failure Cases

My3DGen struggles to reconstruct objects that obscure the face, such as hands and phones, even with PTI. The cause for this is a deficiency in corresponding images of objects in the pre-training facial dataset, FFHQ. Further works may design an EG3D-specific encoder that can encode objects into the latent space similar to Live3DPortrait [70].

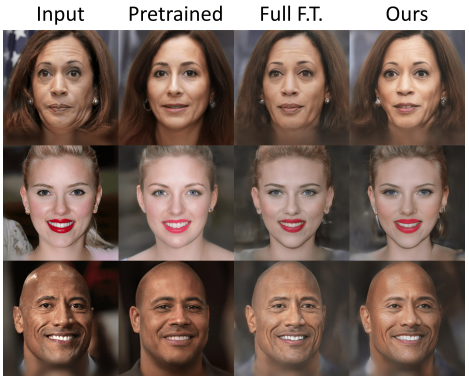

Fig. 12: Image inversion results without optimizing network weights. F.T. indicates fine-tuning and ours is My3DGen.

Fig. 13: Cases where the inversion method fails to reconstruct objects.