

DiffusionRig: Learning Personalized Priors for Facial Appearance Editing

Zheng Ding^{1†}, Xuaner Zhang², Zhihao Xia², Lars Jebe², Zhuowen Tu¹, Xiuming Zhang²
¹UC San Diego ²Adobe

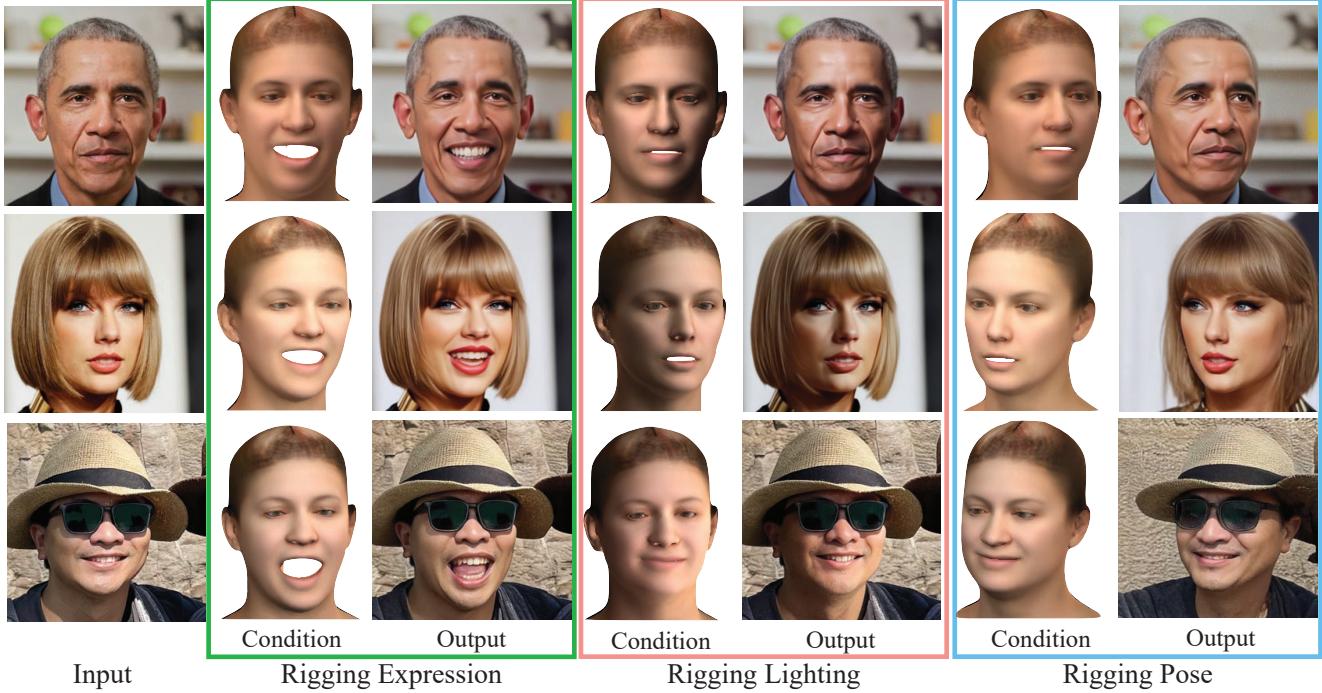


Figure 1. DiffusionRig takes in coarse physical rendering as the condition to “rig” the input image with learned personal priors. The edited images respect the rendering conditions, preserve the identity, and exhibit high-frequency facial details.

Abstract

We address the problem of learning person-specific facial priors from a small number (e.g., 20) of portrait photos of the same person. This enables us to edit this specific person’s facial appearance, such as expression and lighting, while preserving their identity and high-frequency facial details. Key to our approach, which we dub DiffusionRig, is a diffusion model conditioned on, or “rigged by,” crude 3D face models estimated from single in-the-wild images by an off-the-shelf estimator. On a high level, DiffusionRig learns to map simplistic renderings of 3D face models to realistic photos of a given person. Specifically, DiffusionRig is trained in two stages: It first learns generic facial priors from a large-scale face dataset and then person-specific priors from a small portrait photo collection of the person of interest. By learning the CGI-to-photo mapping with such personalized priors,

DiffusionRig can “rig” the lighting, facial expression, head pose, etc. of a portrait photo, conditioned only on coarse 3D models while preserving this person’s identity and other high-frequency characteristics. Qualitative and quantitative experiments show that DiffusionRig outperforms existing approaches in both identity preservation and photorealism. Please see the project website: <https://diffusionrig.github.io> for the supplemental material, video, code, and data.

1. Introduction

It is a longstanding problem in computer vision and graphics to photorealistically change the lighting, expression, head pose, etc. of a portrait photo while preserving the person’s identity and high-frequency facial characteristics. The difficulty of this problem stems from its fundamentally under-constrained nature, and prior work typically addresses this with zero-shot learning, where neural networks were trained on a large-scale dataset of different identities and tested on a new identity. These methods ignore the fact that such

† Work done during an internship at Adobe.

generic facial priors often fail to capture the test identity’s high-frequency facial characteristics, and multiple photos of the same person are often readily available in the person’s personal photo albums, e.g., on a mobile phone. In this work, we demonstrate that one can convincingly edit a person’s facial appearance, such as lighting, expression, and head pose, while preserving their identity and other high-frequency facial details. Our key insight is that we can first learn generic facial priors from a large-scale face dataset [19] and then finetune these generic priors into personalized ones using around 20 photos capturing the test identity.

When it comes to facial appearance editing, the natural question is what representation one uses to change lighting, expression, head pose, hairstyle, accessories, etc. Off-the-shelf 3D face estimators such as DECA [9] can already extract, from an in-the-wild image, a parametric 3D face model that comprises parameters for lighting (spherical harmonics), expression, and head pose. However, directly rendering these physical properties back into images yields CGI-looking results, as shown in the output columns of Figure 1. The reasons are at least three-fold: (a) The 3D face shape estimated is coarse, with mismatched face contours and misses high-frequency geometric details, (b) the assumptions on reflectance (Lambertian) and lighting (spherical harmonics) are restrictive and insufficient for reproducing the reality, and (c) 3D morphable models (3DMMs) simply cannot model all appearance aspects including hairstyle and accessories. Nonetheless, such 3DMMs provide us with a useful representation that is amenable to “appearance rigging” since we can modify the facial expression and head pose by simply changing the 3DMM parameters as well as lighting by varying the spherical harmonics (SH) coefficients.

On the other hand, diffusion models [15] have recently gained popularity as an alternative to Generative Adversarial Networks (GANs) [11] for image generation. Diff-AE [33] further shows that when trained on the autoencoding task, diffusion models can provide a latent space for appearance editing. In addition, diffusion models are able to map pixel-aligned features (such as noise maps in the vanilla diffusion model) to photorealistic images. Although Diff-AE is capable of interpolating from, e.g., smile to no smile, after semantic labels are used to find the direction to move towards, it is unable to perform edits that require 3D understanding and that cannot be expressed by simple binary semantic labels. Such 3D edits, including relighting and head pose change, are the focus of our work.

To combine the best of both worlds, we propose DiffusionRig, a model that allows us to edit or “rig” the appearance (such as lighting and head pose) of a 3DMM and then produce a photorealistic edited image conditioned on our 3D edits. Specifically, DiffusionRig first extracts rough physical properties from single portrait photos using an off-the-shelf method [9], performs desired 3D edits in the 3DMM space,

and finally uses a diffusion model [15] to map the edited “physical buffers” (surface normals, albedo, and Lambertian rendering) to photorealistic images. Since the edited images should preserve the identity and high-frequency facial characteristics, we first train DiffusionRig on the CelebA dataset [27] to learn generic facial priors so that DiffusionRig knows how to map surface normals and the Lambertian rendering to a photorealistic image. Note that because the physical buffers are coarse and do not contain sufficient identity information, this “Stage 1 model” provides no guarantee for identity preservation. At the second stage, we finetune DiffusionRig on a tiny dataset of roughly 20 images of one person of interest, producing a person-specific diffusion model mapping physical buffers to photos of just this person. As discussed, there are appearance aspects not modeled by the 3DMM, including but not limited to hairstyle and accessories. To provide our model with this additional information, we add an encoder branch that encodes the input image into a global latent code (“global” in contrast to physical buffers that are pixel-aligned with the output image and hence “local”). This code is chosen to be low-dimensional in the hope of capturing just the aspects *not* modeled by the 3DMM, such as hairstyle and eyeglasses.

In summary, our contributions are:

- A deep learning model for 3D facial appearance editing (that modifies lighting, facial expression, head pose, etc.) trained using just images with no 3D label,
- A method to drive portrait photo generation using diffusion models with 3D morphable face models, and
- A two-stage training strategy that learns personalized facial priors on top of generic face priors, enabling editing that preserves identity and high-frequency details.

2. Related Work

Our work is related to generative models, 3D Morphable Face Models (3DMMs), and personalized priors.

Generative Modeling Since the proposal of early Generative Adversarial Networks (GANs) [11], researchers have made significant progress in generating photorealistic images of constrained classes, such as faces [18, 20, 21]. Recently, denoising diffusion models [16], which learn to denoise random noise images into photorealistic images, have shown impressive synthesis results and gained popularity as an alternative to GANs. Different diffusion models are invented for faster sampling [41] (used in this work), conditional generation [7, 31], and later pixel-aligned conditional generation [37]. Similarly, we use pixel-aligned conditions, specifically surface normal, albedo, and Lambertian rendering images, as the condition that our diffusion model should satisfy. Closely related to DiffusionRig are Diffusion Autoencoders (Diff-AE) that learn a latent space of facial attributes (e.g., +smiling vs. -smiling) via the autoencoding task [34]. Given binary labels of a certain attribute, the authors find the



Figure 2. **Reconstruction with vs. without personalized priors.** Given the input image and its conditions (surface normals, albedo, and Lambertian rendering) automatically extracted using DECA, Stage 1 learns only generic face priors and fails to reconstruct the identity in both of the randomly sampled reconstructions. With Stage 2, DiffusionRig is able to faithfully reconstruct the input image using either of the two stochastically sampled noise maps.

direction, along which the latent code should be pushed, to manipulate that attribute. 3D-aware generative models are a recent popular trend to combine 3D controllability with 2D image generation [3, 4, 10, 13, 43, 45, 46, 54].

Facial Appearance Modeling 3D Morphable Face Models or 3DMMs provide a valuable parameter space to describe (and in turn solve for) 3D facial characteristics [2]. The FLAME face model learned from 4D scans is a widely-used 3DMM that supports shape, pose, and expression change [24]. We refer the reader to a recent survey paper on Morphable Face Models [8].

RingNet regresses FLAME parameters from 2D images [38]. Also a learning-based method, DECA additionally predicts albedo and lighting in spherical harmonics (SH) from a single face image [9]. An alternative to using 3DMMs for “face de-rendering” is directly predicting surface normals, albedo, and lighting in the image space, as in SfSNet [39]. Although such approaches enjoy the benefit of being able to represent hair, accessories, etc., image-space representations do not provide a physically meaningful parameter space for rigging like 3DMMs do.

The geometry, albedo, and lighting from 3DMM are still extremely coarse and far from reality. The community has bridged the realism gap between 3DMM rendering and real photos through expensive hardware setups to capture fine-grained facial geometry [48, 50] and reflectance fields [6]. Neural network-based, implicit appearance models have also been proposed to address the infeasibility of explicitly describing the appearance with precise reflectance and lighting [1, 12, 28, 29, 30, 35, 36, 40, 42, 52, 53].

Personalized Priors Learning personal priors has been more widely discussed in super-resolution, face restoration, and inpainting, by using exemplar imagery [47], personal supplemental attributes [51], an attention module with identity penalty [49], or facial component dictionaries [25]. Conditional portrait image editing also shares the objective of preserving the input identity [26, 44].

However, it remains a challenge how to compute an unbiased identity score, and these approaches do not explicitly learn personalized priors.

Closer to DiffusionRig that learns a personal prior from a set of personal album of the person, MyStyle [32] is a method to finetune a pre-trained StyleGAN model to achieve a generative model for a specific identity, while preserving the expressiveness of the latent space. However, it does not support precise 3D rigging to control the generation and requires a much larger personal dataset to obtain a smooth personalized latent space. DiffusionRig, on the other hand, focuses on controllable image editing and achieves the smooth editing naturally with the continuous physical space as conditions.

3. Preliminaries

We provide the background knowledge for the two building blocks—3D Morphable Face Models (3DMMs) and Denoising Diffusion Probabilistic Models (DDPMs).

3.1. 3D Morphable Face Models

3D Morphable Face Models or 3DMMs are parametric models that use a compact latent space (handcrafted or learned from scans) to represent the head pose, face geometry, facial expression, etc. [2, 24]. In this paper, we employ FLAME [24], a popular 3DMM using standard vertex-based linear blend skinning with corrective blendshapes and representing a face mesh with pose, shape, and expression parameters. Although FLAME provides a compact and physically meaningful space for face *geometry*, it does not provide descriptions for *appearance*. To this end, DECA [9] uses FLAME and additionally models facial appearance with Lambertian reflectance and Spherical Harmonics (SH) lighting. Trained on a large dataset, DECA predicts albedo, SH lighting, and the FLAME parameters from a single portrait image. We utilize DECA to generate *rough* 3D representations that support easy “rigging” by editing the FLAME parameters, albedo, and/or SH lighting. As we show in Figure 2, the realism gap between DECA rendering and real photos is significant, calling for measurements post-editing.

3.2. Denoising Diffusion Probabilistic Models

Denoising Diffusion Probabilistic Models (DDPMs) [7, 15, 31] are a class of generative models that take random

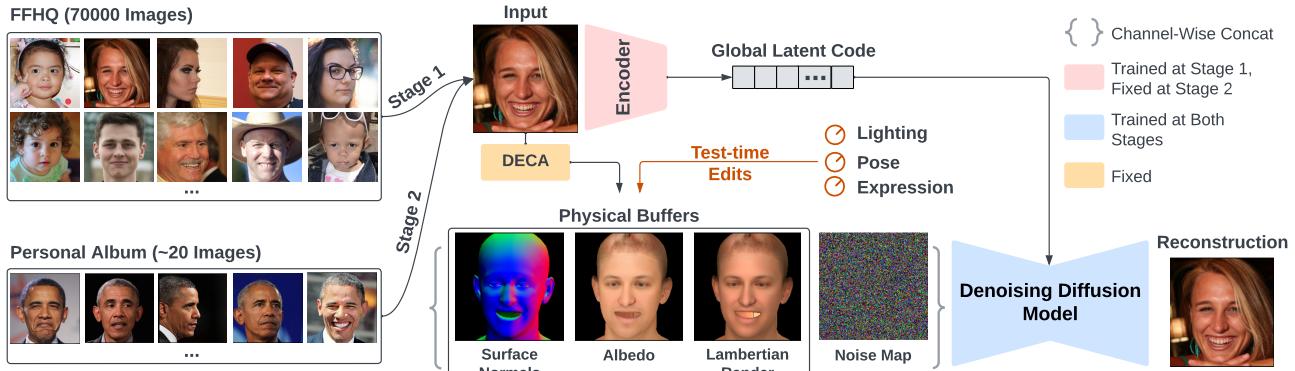


Figure 3. **DiffusionRig** overview. The input to our model is a set of physical buffers reconstructed from the input, a random noise map, and a global latent code that encodes nuance features not modeled by the physical buffers. At Stage 1, we train our model on a large face dataset to learn generic face priors. At Stage 2, we keep the global latent code encoder frozen and fine-tune the diffusion model to learn personalized priors.

noise images as input and denoise the images progressively to produce photorealistic images. This generation process can be seen as the reverse of the diffusion process that gradually adds noise to images. The key component of DDPMs is a denoising network f_θ . During training, it takes a noisy image x_t and a timestep t ($1 \leq t \leq T$), and predicts the noise at time t : $\hat{\epsilon}_t = f_\theta(x_t, t)$, where $x_t = \alpha_t x_0 + \sqrt{1 - \alpha_t^2} \epsilon_t$, ϵ_t is a random, normally distributed noise image, and α_t is a hyperparameter that gradually increases noise level of x_t with each step of the forward process. The loss is computed on the distance between ϵ_t and $\hat{\epsilon}_t$. Therefore, the trained model can generate images by taking as input a random noise image and gradually denoising it to a photorealistic one.

4. Method

To enable personalized appearance editing, our model, which we dub DiffusionRig, needs to (a) generate images based on different appearance conditions, such as novel lighting, and (b) learn personal priors so that the person’s identity is not altered during editing.

To this end, we design a two-stage training pipeline as shown in Figure 3. At the first stage, the model learns generic face priors by being trained to reconstruct portrait images given their underlying “appearance conditions” represented as physical buffers automatically extracted using an off-the-shelf estimator. At the second stage, we finetune our model using portrait photos of just one person so that the model learns personalized priors, which are necessary to prevent identity shift during appearance editing.

4.1. Learning Generic Face Priors

Our first stage is designed to learn facial priors that enable photorealistic image synthesis conditioned on physical constraints like lighting. For the physical conditioning, we use DECA [9] to produce the physical parameters including

the FLAME [24] parameters (shape β , expression ψ , and pose θ), albedo α , (orthographic) camera \mathbf{c} , and (spherical harmonics) lighting \mathbf{l} from the input portrait image. We then use the Lambertian reflectance to render these physical properties into three buffers: surface normals, albedo, and Lambertian rendering. Although these physical buffers provide pixel-aligned descriptions of the facial geometry, albedo, and lighting, they are rather coarse and nowhere close to photorealistic images (see the Lambertian rendering in Figures 1 and 2). Still, using these buffers, we can “rig” our generative model in a disentangled, physically meaningful way by changing the DECA parameters. For photorealistic image synthesis, we use a Denoising Diffusion Probabilistic Model (DDPM) as our generator because DDPMs can naturally take pixel-aligned conditions (more advantageous than latent code conditions as shown in Section 5.5) to drive the generation process.

Besides the pixel-aligned physical buffers, we keep the random noise images in DDPMs to explain the stochasticity during generation. In addition to the pixel-aligned buffers and noise map, we need another condition to encode *global* appearance information (as opposed to local information such as local surface normals) that is not modeled by the physical buffers, such as hair, hat, glasses, and the image background. Therefore, our diffusion model takes both physical buffers and a learned global latent code as conditions for image synthesis. Formally, our model can be described as $\hat{\epsilon}_t = f_\theta([x_t, z], t, \phi_\theta(x_0))$ where x_t is the noisy image at timestep t , z represent the physical buffers, x_0 is the original image, $\hat{\epsilon}_t$ is the predicted noise, and f_θ and ϕ_θ are the denoising model and the global latent encoder, respectively.

It is theoretically possible that the global latent code also encodes local geometry, albedo, and/or illumination information, which could lead to the diffusion model ignoring the physical buffers entirely. Empirically, we find that the network learns to use the physical buffers for local information

and does not rely on the global latent code, possibly because these buffers are pixel-aligned with the ground truth and thus more easily leveraged by the model.

4.2. Learning Personalized Priors

After learning the generic facial priors at the first stage, DiffusionRig is able to generate photorealistic images given coarse physical buffers. The next step is to learn personalized priors for a given person to avoid identity shift during appearance editing. Personal priors are crucial to preserving identity and high-frequency facial characteristics, as shown in Figure 2. We achieve this by finetuning our denoising model on a specific person’s photo album of around 20 images. During the finetuning stage, the denoising model learns the person’s identity information. We fix the global encoder from the previous stage since it has learned to encode global image information not modeled by the physical buffers (which we want to retain). We show that this approach is simple and yet effective compared with GANs that need careful tuning, as mentioned in MyStyle [32].

For this small personalized dataset, we also extract the DECA parameters first. However, since DECA is a single-image estimator, its output is sensitive to extreme poses or expressions. Under the assumption that the general shape of a person’s face does not change drastically within a reasonable period of time, we compute the mean of the shape parameters in FLAME over all the images in the album and use that mean shape when conditioning DiffusionRig.

4.3. Model Architecture

DiffusionRig consists of two trainable parts: a denoising model f_θ and a global encoder ϕ_θ . The architecture of our denoising model is based on ADM [7] with modifications to reduce computational cost and take an additional global latent code as input. For the global code, we use the same method that ADM uses for their time embedding: We scale and shift the features in each layer using the global latent code. The encoder is simply a ResNet-18 [14] and we use the output features as the global latent codes.

Our loss function is a P2 weight loss [5] that computes distances between predicted and ground-truth noises: $\mathcal{L} = \lambda'_t \|\hat{\epsilon}_t - \epsilon_t\|_2^2$, where λ'_t is a hyperparameter to control the loss weight at different timesteps. We empirically find that the P2 weight loss speeds up the training process and generates high-quality images compared with a constant loss weight.

4.4. Implementation Details

During the first stage, we train DiffusionRig on the FFHQ dataset [19], which contains 70,000 images. With Adam [23] as the optimizer with a learning rate of 10^{-4} , we train DiffusionRig for 50,000 iterations with a batch size of 256 (so the total number of samples seen by the model is 12,800,000). During the second stage, we use only 10–20 images of a single person. In the following, we show results for four

celebrities (Obama, Biden, Swift, and Harris) and two non-celebrities. Please see the supplemental material and video for more results including more identities. We use 20 images for each person except for Harris, for whom we use only 10, and for the ablation study on the number of training images. We provide the personal photo album of two identities in the supplemental material. We finetune our model on each small dataset for 5,000 iterations with a batch size of 4 (so the total number of seen samples during finetuning is 20,000). We furthermore decrease the learning rate to 10^{-5} for the second stage. Training for the first stage takes around 15 hours using eight A100 GPUs, and the Stage 2 finetuning completes within 30 minutes on a single V100 GPU.

5. Experiments

We first show how to edit a person’s appearance (e.g., facial expression, lighting, and head pose) by modifying the physical buffers that condition the model. We then demonstrate how to rig, with the global latent code, other aspects of a person’s appearance not modeled by the physical buffers such as hairstyle and accessories. By swapping in the global latent code from another image, we can transfer portrait characteristics, such as hairstyle, accessories including glasses, and/or the image background, while preserving the physical properties (e.g., identity, pose, expression, and lighting) from the original image. Finally, we show the power of the learned personal priors by conditioning, for example, an Obama model on both the physical buffers and global latent code from a different person (to “Obama-fy” that person).

5.1. Rigging Appearance With Physical Buffers

In this section, we use our personalized model to rig the appearance with physical buffers. We show three different types of appearance rigging: relighting, expression change, and pose change. For relighting, we use different Spherical Harmonics (SH) parameters for producing the Lambertian rendering. To change the expression, we modify the expression and jaw rotation parameters of FLAME (the last three parameters of the pose vector). To vary the pose, we modify the head rotation parameters (the first three parameters of the pose vector). The 64-dimensional global latent code is produced by encoding the input image and remains unchanged when editing appearance.

Our results are displayed in Figure 4, where we depict three identities: two celebrities and one daily user. All the images have a resolution of 256×256 . Additionally, 512×512 results can be found in the supplemental material. We compare our method against DECA [9], HeadNerf [17], GIF [10], and MyStyle [32], of which the first two are 3D face model estimation methods, and the latter two are GAN-based approaches. As Figure 4 shows, while GIF is capable of rigging the appearance by changing the expression and pose, it fails to preserve the individual’s identity. DiffusionRig and MyStyle, on the other hand, are both personalized models

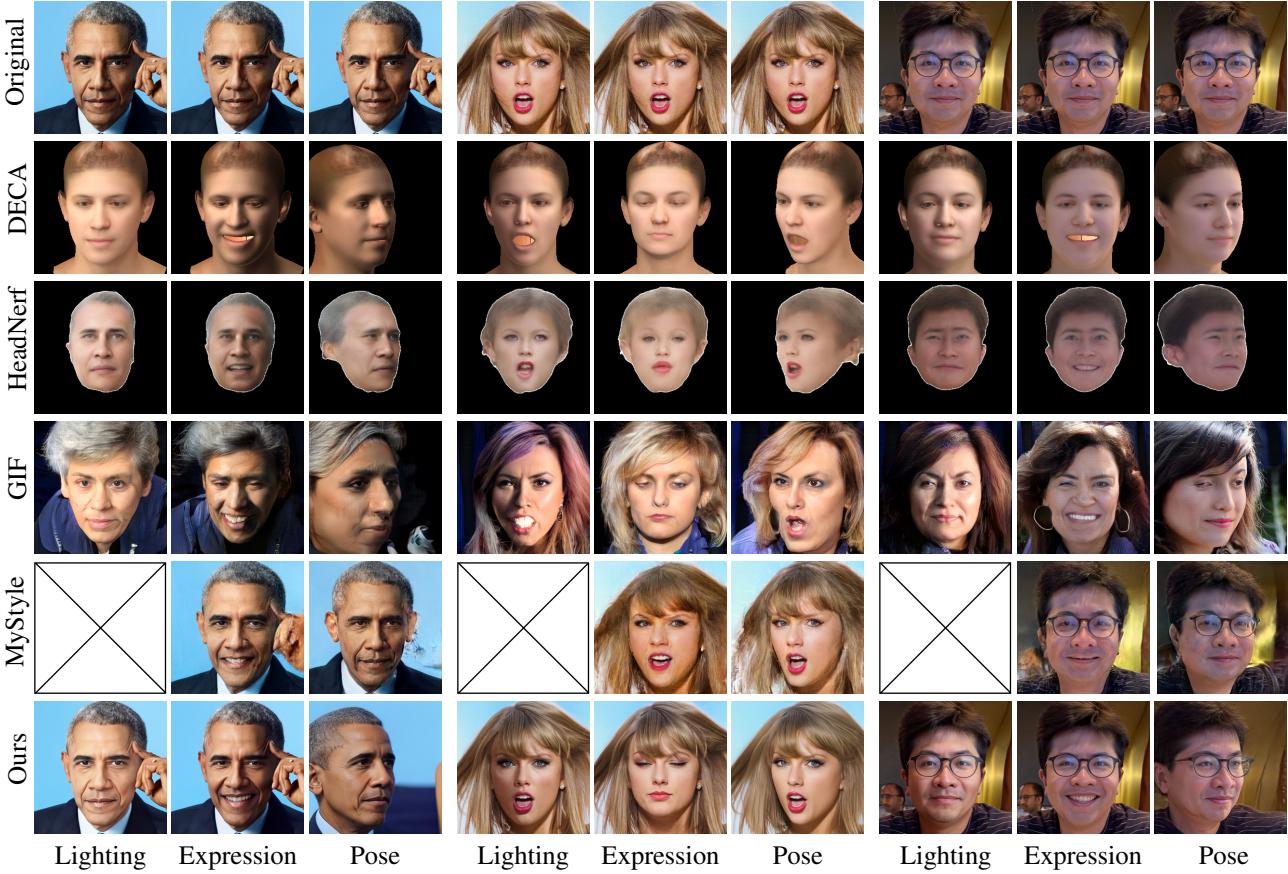


Figure 4. **Appearance editing.** DiffusionRig achieves convincing appearance edits while preserving the individual’s identity using only 20 images per identity. GIF creates realistic-looking images but does not use personalized priors, leading to significant identity shifts. MyStyle is unable to make dramatic changes to the expression or pose without artifacts or minor identity shifts. In addition, MyStyle does not trivially support controllable relighting, so the corresponding fields have been left empty.

that are able to preserve the identity. However, since our method is directly conditioned on physical buffers, we can rig the appearance in a physically-based manner, whereas MyStyle needs to search for and step into a certain direction within the latent space to produce the target appearance, limiting its controllability, interpretability, and capacity for dramatic appearance changes. We also observe more artifacts for MyStyle when doing appearance editing, which is likely due to the use of too few images during finetuning the StyleGAN model.

5.2. Rigging Appearance With Global Latent Code

By design, DiffusionRig finds it easier to learn what physical buffers can describe from the pixel-aligned buffers than from the global latent code. The latent code thus encodes what physical buffers cannot describe including background, makeup, and hairstyle. In this part, we change the global latent code to show its effects on the generated images.

In Figure 5, we show a 2×3 matrix of generated images. Along the horizontal axis, we swap in the global latent code from another image of the same person while keeping the physical buffers identical (i.e., same physical buffers but

different global codes). Along the vertical axis, we replace the physical buffers while keeping the same global latent code (i.e., same global code but different physical buffers). We can see that geometry information, such as head pose and expression, is preserved for each row, which shows that only the physical buffers (not the latent code) contain such information. This means that in DiffusionRig these physical properties are well disentangled from each other and from other appearance properties that physical buffers cannot describe. On the other hand, the information hard to model explicitly, including image background, glasses, and hair style/color, is encoded in the global latent code.

5.3. Identity Transfer With Learned Priors

In previous sections, we saw what information the physical buffers and the global latent code encode. Now, we demonstrate what information is encoded in the personalized diffusion models’ weights. Here, we keep both physical buffers and global latent code the same but exchange the personalized model itself with another person’s personalized model (i.e., model swapping without code or buffer swapping). The results of this experiment for four identities are



Figure 5. Mix and match of physical buffers and global latent code. We mix the physical buffers from one image and the global latent code from another image to demonstrate how the two conditions encode disentangled information.

shown in Figure 6. Each row uses the same physical buffers and latent code but another personalized model. Each column uses the same personalized model but different physical buffers and latent code. For example, the column “Obama-fy” shows four images that are generated by Obama’s personal model but using the other celebrities’ images as input. We see that across each row, while all inputs (physical buffers plus global latent code) are the same, the four different personalized models output different identities. These results further corroborate that our model is able to learn personalized priors from a small dataset.

5.4. Baseline Comparisons & Evaluation Metrics

We evaluate our DiffusionRig quantitatively in three aspects: rigging quality, identity preservation, and photorealism, since these three qualities are the most important for our personalized appearance editing.

DECA Re-Inference Error We follow the same setup as in GIF [10] to compute the DECA re-inference error. To evaluate relighting quality, we directly compute the RMSE on the re-inferred spherical harmonics. We show our results in Table 1. For our model, we also evaluate two ablated versions: “vector cond.” and “feature cond.” Instead of using pixel-aligned physical buffers as the condition, we use DECA’s output parameters and features computed from physical buffers as conditions in our two ablated models. More details can be found in Section 5.5.

Face Re-Identification Error An important metric for evaluating this work is whether DiffusionRig can preserve the identity after appearance editing, since identity shift is a notorious problem in generative model-based editing. To this end, we run a widely popular face re-identification network [22] to automatically determine if the edited and



Figure 6. Swapping personalized models. We demonstrate the power of personalized priors by running one person’s model on other identities. This creates the effect of “adding” one person’s identity to another person. The images with green borders are “no-swap” results where the corresponding person’s model is used.

	Light ↓	Shape ↓	Exp. ↓	Pose ↓
GIF [10]	13.8	3.0	5.0	5.6
GIF, vector cond. [10]	–	3.4	23.1	29.7
DiffusionRig (Ours)	11.2	4.3	2.8	4.2
DiffusionRig, vector cond.	15.5	10.7	8.8	14.0
DiffusionRig, feature cond.	27.0	5.3	4.1	21.6

Table 1. RMSE of DECA re-inference. All numbers are multiplies of 10^{-3} . We generate 1,000 images for evaluation. For shape, expression, and pose, the RMSE is computed on rendered FLAME faces. For lighting, the RMSE is computed on re-inferred spherical harmonics directly. We only use our Stage 1 model since GIF is not a personalized model. Numbers for GIF and its vector-conditioned variant are cited from the original paper [10].

original images are of the same person. As Table 2 shows, both MyStyle [32] and DiffusionRig preserve the identity in all 400 expression-edited images of Obama and another 400 of Swift. That said, for dramatic changes such as head pose change, DiffusionRig preserves the identity better than MyStyle, as also demonstrated by Figure 4. One caveat of this error metric, though, is the obvious degenerate solution of not applying any edit at all, thereby achieving a perfect score. We refer the reader to Figure 4 and Table 1, which show that DiffusionRig avoids this degenerate solution.

User Study To further evaluate both the photorealism and identity preservation of images from DiffusionRig against MyStyle, we conduct a user study involving Amazon Mechanical Turk. During the study, we show pairs of images,

	Auto. Face Re-ID ↑		User Study ↑							
	Obama and Swift		Obama		Swift		Obama		Pose	
MyStyle	Expr. 100%	Pose 97.9%	Expr. 79.4%	Pose 78.0%	Expr. 64.5%	Pose 62.5%	Expr. 87.2%	Pose 86.5%	Expr. 82.4%	Pose 80.2%
Ours	100%	99.3%	87.2%	86.5%	82.4%	80.2%				

Table 2. **DiffusionRig vs. MyStyle [32]** in expression and pose editing, as measured by an automatic face re-ID error [22] (which has an obvious flaw; see text) as well as a user study on both realism and identity preservation.

where the left image is an original image from the real image dataset, and the right image is a generated one. We occasionally include some real images on the right, too, for consistency check and quality control. We then ask the users whether the right image is a real image of the person on the left (so both photorealism and identity preservation are probed). We generate images that include either an expression or pose change for both DiffusionRig and MyStyle. We report our results in Table 2.

5.5. Ablation Study

We show several ablation studies to motivate the finetuning stage that injects the personalized prior and the choice of physical, pixel-aligned buffers to condition the model.

No Personalized Priors We first show how DiffusionRig performs in the absence of personalized priors (i.e., trained on only the large dataset from Stage 1). Figure 2 shows that our model learns to use the physical buffers as conditions for pose, expression, and lighting, but it is incapable of preserving the person’s identity during appearance editing.

Number of Images Here we explore how the number of images used in Stage 2 affects DiffusionRig’s ability of learning personalized priors. We train three models of a non-celebrity with 1, 5, 10, and 20 images and test them on relighting, expression change, and pose change. As Figure 7 demonstrates, using just 1, 5, or 10 images yields worse results than using 20 images (unsurprisingly). With more images, DiffusionRig learns better-personalized priors that capture high-frequency face characteristics, such as the wrinkles in Figure 7.

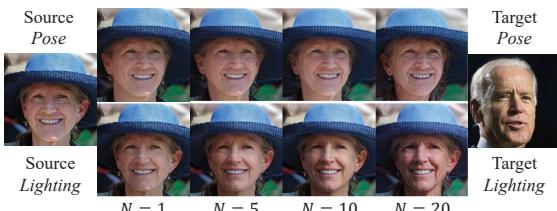


Figure 7. **Quality w.r.t. number of Stage 2 images.** DiffusionRig achieves high-quality relighting and pose change with 20 images for Stage 2. Using fewer may yield blurry results and make them hard to rig with new conditions.

Different Forms of Conditions There are alternative ways to condition the image synthesis. We demonstrate that

pixel-aligned physical buffers are the most effective form in accurately rigging the appearance. We explore the following two conditioning alternatives. “**Vector cond.**” is when we directly concatenate DECA parameters, a 236-dimensional vector, to the global latent code without using pixel-aligned buffers. “**Feature cond.**” means that we concatenate the physical buffers to the input image and pass them into the encoder to compute a global latent code, which is then used as a non-spatial feature condition. As shown in Figure 8, using pixel-aligned physical guidance is essential for accurate conditional image editing. Both vector and feature conditioning suffer from the generated images not following the desired physical guidance.

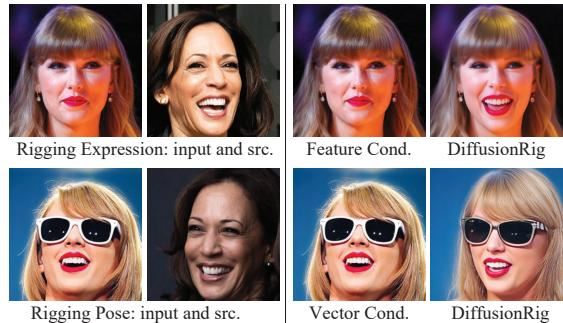


Figure 8. **Ablation on the form of conditions.** Neither feature conditioning nor vector conditioning is able to rig the input image to follow the physical properties of the target image.

6. Limitations & Conclusion

Although DiffusionRig achieves state-of-the-art facial appearance editing, it relies on a small portrait dataset to finetune, which limits its scalability for massive user adoption. Furthermore, when the edit involves dramatic head pose change, DiffusionRig may not stay faithful to the original background, since head pose change sometimes reveals what used to be occluded, therefore requiring background inpainting—a topic beyond the scope of this paper. Additionally, since DiffusionRig relies on DECA to get physical buffers, it will also be affected by DECA’s limited estimation capability: for instance, extreme expressions usually cannot be well predicted, and the estimated lighting is sometimes coupled with the skin tone.

In this paper, we have presented DiffusionRig, a riggable diffusion model for identity-preserving, personalized editing of facial appearance. We introduced a two-stage method to first learn generic face priors and later personalized priors. Using both explicit conditioning via physical buffers and implicit conditioning via global latent code, we can drive and control our model’s facial image synthesis.

Acknowledgment We thank Marc Levoy for the valuable feedback and everyone whose photos appear in this paper for their permission.

References

- [1] Sai Bi, Stephen Lombardi, Shunsuke Saito, Tomas Simon, Shih-En Wei, Kevyn Mcphail, Ravi Ramamoorthi, Yaser Sheikh, and Jason Saragih. Deep relightable appearance models for animatable faces. 40(4):15. 3
- [2] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999. 3
- [3] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3d generative adversarial networks. (*arXiv:2112.07945*), Apr 2022. arXiv:2112.07945 [cs]. 3
- [4] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. 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*, pages 5799–5809, 2021. 3
- [5] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11472–11481, 2022. 5, A2
- [6] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 145–156, 2000. 3
- [7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 2, 3, 5, A1, A2
- [8] Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 3d morphable face models – past, present and future. *arXiv:1909.01815 [cs]*, Apr 2020. arXiv: 1909.01815. 3
- [9] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021. 2, 3, 4, 5
- [10] Partha Ghosh, Pravir Singh Gupta, Roy Uziel, Anurag Ranjan, Michael J Black, and Timo Bolkart. Gif: Generative interpretable faces. In *2020 International Conference on 3D Vision (3DV)*, pages 868–878. IEEE, 2020. 3, 5, 7
- [11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2
- [12] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18653–18664, 2022. 3
- [13] Zekun Hao, Arun Mallya, Serge Belongie, and Ming-Yu Liu. Gancraft: Unsupervised 3d neural rendering of minecraft worlds. *arXiv:2104.07659 [cs]*, Apr 2021. arXiv: 2104.07659. 3
- [14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5
- [15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 3
- [16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, page 6840–6851. Curran Associates, Inc., 2020. 2
- [17] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 5
- [18] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021. 2
- [19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2, 5
- [20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. (*arXiv:1812.04948*), Mar 2019. arXiv:1812.04948 [cs, stat]. 2
- [21] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. (*arXiv:1912.04958*), Mar 2020. arXiv:1912.04958 [cs, eess, stat]. 2
- [22] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009. 7, 8
- [23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5
- [24] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 3, 4
- [25] Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration via deep multi-scale component dictionaries. In *European Conference on Computer Vision*, pages 399–415. Springer, 2020. 3
- [26] Yuchen Liu, Zhixin Shu, Yijun Li, Zhe Lin, Richard Zhang, and SY Kung. 3d-fm gan: Towards 3d-controllable face manipulation. In *European Conference on Computer Vision*, pages 107–125. Springer, 2022. 3
- [27] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 2
- [28] Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. Deep appearance models for face rendering. *ACM Transactions on Graphics*, 37(4):1–13, Jul 2018. arXiv: 1808.00362. 3
- [29] Abhimitra Meka, Christian Haene, Rohit Pandey, Michael

- Zollhöfer, Sean Fanello, Graham Fyffe, Adarsh Kowdle, Xueming Yu, Jay Busch, Jason Dourgarian, et al. Deep reflectance fields: high-quality facial reflectance field inference from color gradient illumination. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 3
- [30] Thomas Nestmeyer, Jean-François Lalonde, Iain Matthews, and Andreas Lehrmann. Learning physics-guided face relighting under directional light. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5124–5133, 2020. 3
- [31] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 2, 3
- [32] Yotam Nitzan, Kfir Aberman, Qiurui He, Orly Liba, Michal Yarom, Yossi Gandelsman, Inbar Mosseri, Yael Pritch, and Daniel Cohen-Or. Mystyle: A personalized generative prior. *arXiv preprint arXiv:2203.17272*, 2022. 3, 5, 7, 8
- [33] Konpat Preechakul, Nattanat Chathee, Suttisak Wizadwongsu, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10619–10629, 2022. 2
- [34] Konpat Preechakul, Nattanat Chathee, Suttisak Wizadwongsu, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. *arXiv:2111.15640 [cs]*, Mar 2022. arXiv: 2111.15640. 2
- [35] Mallikarjun B. R, Ayush Tewari, Abdallah Dib, Tim Weyrich, Bernd Bickel, Hans-Peter Seidel, Hanspeter Pfister, Wojciech Matusik, Louis Chevallier, Mohamed Elgharib, and Christian Theobalt. Photoapp: Photorealistic appearance editing of head portraits. *arXiv:2103.07658 [cs]*, Mar 2021. arXiv: 2103.07658. 3
- [36] Edoardo Remelli, Timur Bagautdinov, Shunsuke Saito, Changlei Wu, Tomas Simon, Shih-En Wei, Kaiwen Guo, Zhe Cao, Fabian Prada, Jason Saragih, et al. Drivable volumetric avatars using texel-aligned features. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 3
- [37] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2
- [38] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J. Black. Learning to regress 3d face shape and expression from an image without 3d supervision. (*arXiv:1905.06817*), May 2019. arXiv:1905.06817 [cs]. 3
- [39] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D. Castillo, and David Jacobs. Sfsnet: Learning shape, reflectance and illuminance of faces in the wild. *arXiv:1712.01261 [cs]*, Apr 2018. arXiv: 1712.01261. 3
- [40] Aliaksandra Shysheya, Egor Zakharov, Kara-Ali Aliev, Renat Bashirov, Egor Burkov, Karim Iskakov, Aleksei Ivakhnenko, Yury Malkov, Igor Pasechnik, Dmitry Ulyanov, Alexander Vakhitov, and Victor Lempitsky. Textured neural avatars. page 11. 3
- [41] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2
- [42] Tiancheng Sun, Jonathan T Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul E Debevec, and Ravi Ramamoorthi. Single image portrait relighting. *ACM Trans. Graph.*, 38(4):79–1, 2019. 3
- [43] Feitong Tan, Sean Fanello, Abhimitra Meka, Sergio Orts-Escalano, Danhang Tang, Rohit Pandey, Jonathan Taylor, Ping Tan, and Yinda Zhang. Volux-gan: A generative model for 3d face synthesis with hdri relighting. (*arXiv:2201.04873*), Jan 2022. number: arXiv:2201.04873 arXiv:2201.04873 [cs]. 3
- [44] Ayush Tewari, Mohamed Elgharib, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. Pie: Portrait image embedding for semantic control. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020. 3
- [45] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6142–6151, 2020. 3
- [46] Ayush Tewari, Xingang Pan, Ohad Fried, Maneesh Agrawala, Christian Theobalt, et al. Disentangled3d: Learning a 3d generative model with disentangled geometry and appearance from monocular images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1516–1525, 2022. 3
- [47] Kaili Wang, Jose Oramas, and Tinne Tuytelaars. Multiple exemplars-based hallucination for face super-resolution and editing. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 3
- [48] Lizhen Wang, Zhiyuan Chen, Tao Yu, Chenguang Ma, Liang Li, and Yebin Liu. Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset. (*arXiv:2203.14057*), May 2022. arXiv:2203.14057 [cs]. 3
- [49] Zhouxia Wang, Jiawei Zhang, Runjian Chen, Wenping Wang, and Ping Luo. Restoreformer: High-quality blind face restoration from undegraded key-value pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17512–17521, 2022. 3
- [50] Cheng-hsin Wu, Ningyuan Zheng, Scott Ardisson, Rohan Bali, Danielle Belko, Eric Brockmeyer, Lucas Evans, Timothy Godisart, Hyowon Ha, Alexander Hypes, Taylor Koska, Steven Krenn, Stephen Lombardi, Xiaomin Luo, Kevyn McPhail, Laura Millerschoen, Michal Perdoch, Mark Pitts, Alexander Richard, Jason Saragih, Junko Saragih, Takaaki Shiratori, Tomas Simon, Matt Stewart, Autumn Trimble, Xinshuo Weng, David Whitewolf, Chenglei Wu, Shou-I. Yu, and Yaser Sheikh. Multiface: A dataset for neural face rendering. (*arXiv:2207.11243*), Jul 2022. arXiv:2207.11243 [cs]. 3
- [51] Xin Yu, Basura Fernando, Richard Hartley, and Fatih Porikli. Super-resolving very low-resolution face images with supplementary attributes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 908–917, 2018. 3
- [52] Mona Zehni, Shaona Ghosh, Krishna Sridhar, and Sethu Raman. Joint learning of portrait intrinsic decomposition and relighting. (*arXiv:2106.15305 [cs]*), Jun 2021. arXiv: 2106.15305. 3
- [53] Xuaner Zhang, Jonathan T Barron, Yun-Ta Tsai, Rohit Pandey, Xiuming Zhang, Ren Ng, and David E Jacobs. Portrait

- shadow manipulation. *ACM Transactions on Graphics (TOG)*, 39(4):78–1, 2020. 3
- [54] Xiaoming Zhao, Fangchang Ma, David Güera, Zhile Ren, Alexander G Schwing, and Alex Colburn. Generative multiplane images: Making a 2d gan 3d-aware. In *European Conference on Computer Vision*, pages 18–35. Springer, 2022.

3

Supplementary Material

A. More Results on Personalized Editing

We provide more results on using physical buffers to rig/drive facial appearance generation in Figure S1.



Figure S1. More results on using physical buffers to rig the facial appearance. The physical buffers (not shown) are used to edit the input images (top row) in terms of facial expression (left column), lighting (middle column), and head pose (right column).

B. Extreme Lighting Editing

During the training of DiffusionRig, we rely on the SH-based lighting model from DECA, which is limited in modeling high-frequency lighting. At inference time, we can use

a different lighting representation that can model directional lighting with cast shadow (through ray casting). We show one such extreme lighting example and another RGB lighting example in Figure S2, for which our model regresses slightly towards less extreme lighting but still produces reasonable results.

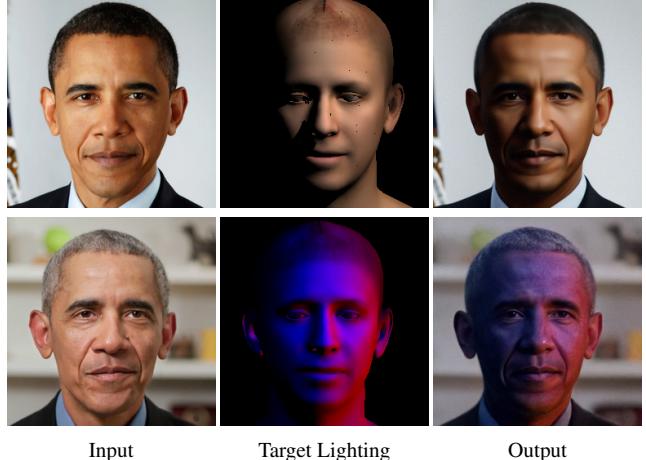


Figure S2. Stress test with difficult directional and RGB lighting.

C. Personal Photo Collections

In Figure S3, we show two sets of images we used to train Stage 2. For celebrities, we crawl the photos from the internet; for non-celebrities, we use everyday photos. In comparison, MyStyle requires 92–279 images for finetuning. When using only 20 images as we do, MyStyle cannot learn personalized priors well as shown in the main paper.

D. Neural Network Architecture Details

For our global encoder, we modify the ResNet-18 model by replacing the final classification layer with a feature extraction layer. More specifically, we change the last layer into a linear layer that outputs our latent code.

Our diffusion model is based on the architecture presented in Guided-Diffusion [7]. We modify the architecture so that the model can take the global latent code as another condition (in addition to the concatenation of physical buffers and the noise image). This global latent code is used for scaling and shifting the features. Our model architecture details can be found in Table S1, where we provide hyperparameters for both our 256×256 and 512×512 models.



Figure S3. Personal photo collections used for training Stage 2: Taylor Swift and a non-celebrity.

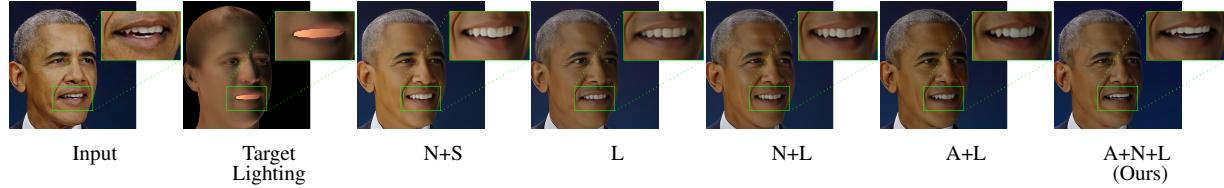


Figure S4. **Ablation on different input conditions.** N: Normals, S: SH layers, L: Lambertian rendering, A: Albedo.

	256×256	512×512
Diffusion Steps	1000	1000
Channels	128	128
Channels Multiple	1, 1, 2, 2, 4, 4	0.5, 1, 1, 2, 2, 4, 4
Heads Channels	128	128
Attention Resolution	16	16
Dropout	0.1	0.1
P2_gamma†	1.0	1.0
P2_k†	1.0	1.0
Optimizer	Adam	Adam
Weight Decay	0.0	0.0
Batch Size (S1)	256	64
Batch Size (S2)	4	2
Iterations (S1)	50k	200k
Iterations (S1)	5k	20k
Learning Rate (S1)	10^{-4}	10^{-4}
Learning Rate (S2)	10^{-5}	10^{-5}

Table S1. **DiffusionRig architecture details.** S1 and S2 denote Stages 1 and 2, respectively. Refer to Guided-Diffusion [7] for more details. † are two hyperparameters defined in prior work [5].

E. Different Types of Pixel-Aligned Buffers

We ablate different pixel-aligned buffers in Figure S4. In our method, we use three kinds of physical buffers from DECA which are Normals (N), Albedo (A) and Lambertian rendering (L). With Lambertian rendering being the only physical buffer that contains lighting information, we include it in all our ablation studies except for the “N+S” where we use Normals and Spherical Harmonics with SH rendered on all-white albedo (i.e., shading), so it doesn’t contain albedo information. We can see with Normals, Albedo, and Lambertian rendering, the results preserve details (e.g., mouth) better, while N+S cannot render accurate lighting due to the missing albedo.

F. Higher-Resolution Results (512×512)

DiffusionRig can be trained at 512×512 resolution. We show these higher-resolution results in Figures S5 and S6 on two new celebrities.

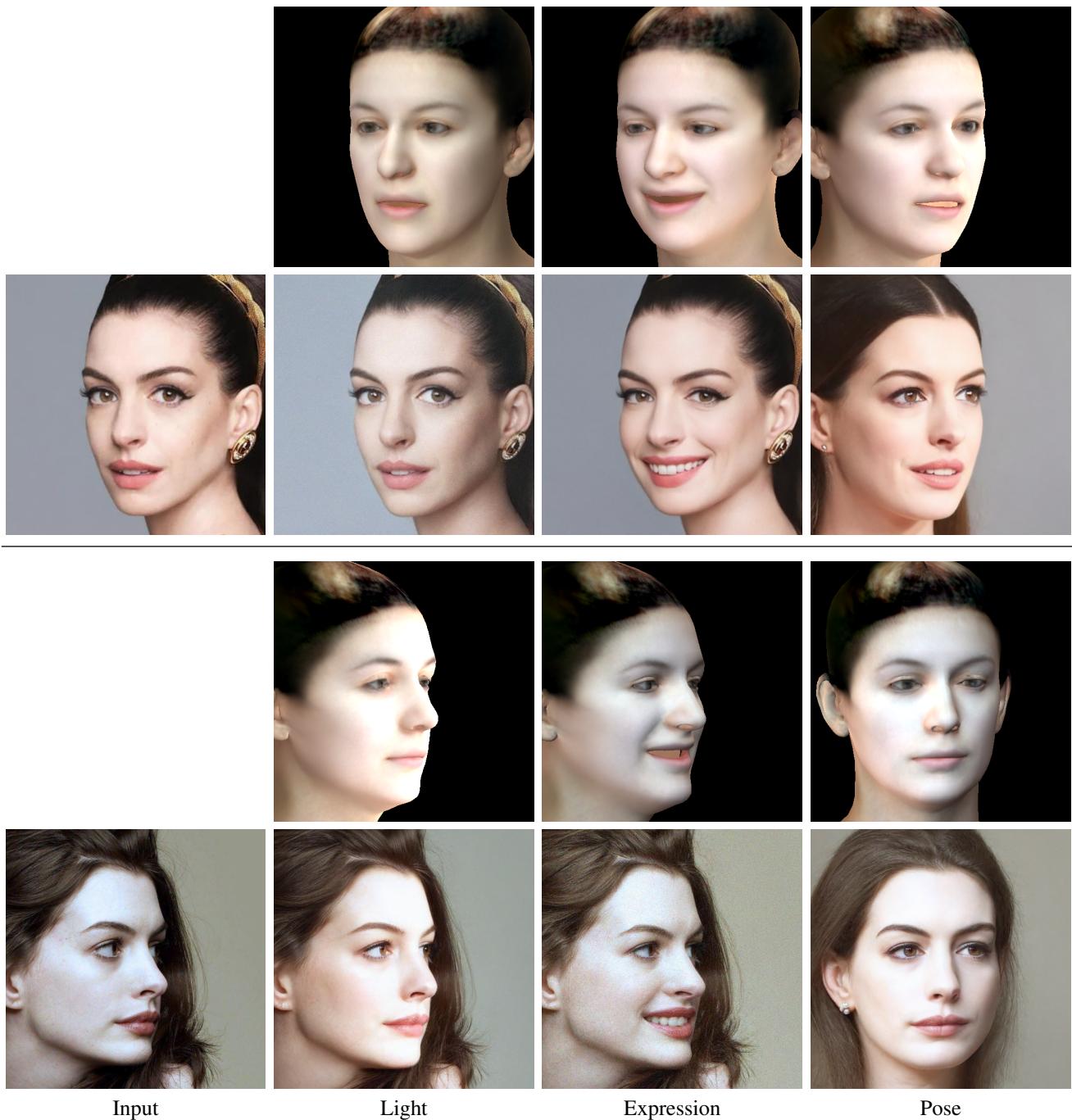


Figure S5. **512×512 Facial Appearance Editing Results.** Two groups of results are presented here with the first row of each group being the physical buffers that drive the editing.



Figure S6. **512×512 Facial Appearance Editing Results.** Two groups of results are presented here with the first row of each group being the physical buffers that drive the editing.