

3D GAN Inversion for Controllable Portrait Image Animation

CONNOR Z. LIN*, DAVID B. LINDELL*, ERIC R. CHAN, and GORDON WETZSTEIN, Stanford University, USA



Fig. 1. Portrait image animation and attribute editing using the proposed technique. Given a source portrait image and a target expression (e.g., specified with a target image), our method transfers the expression and pose to the input source image. We achieve multi-view consistent edits of pose by embedding the expression-edited portrait image into the latent space of a 3D GAN (see predicted underlying shape, right). By interpolating within the latent space of the GAN, we can also apply our method to animate attribute-edited images, allowing adjustments to age, hair, gender or appearance in addition to expression and pose.

Millions of images of human faces are captured every single day; but these photographs portray the likeness of an individual with a fixed pose, expression, and appearance. Portrait image animation enables the post-capture adjustment of these attributes from a single image while maintaining a photorealistic reconstruction of the subject’s likeness or identity. Still, current methods for portrait image animation are typically based on 2D warping operations or manipulations of a 2D generative adversarial network (GAN) and lack explicit mechanisms to enforce multi-view consistency. Thus these methods may significantly alter the identity of the subject, especially when the viewpoint relative to the camera is changed. In this work, we leverage newly developed 3D GANs, which allow explicit control over the pose of the image subject with multi-view consistency. We propose a supervision strategy to flexibly manipulate expressions with 3D morphable models, and we show that the proposed method also supports editing appearance attributes, such as age or hairstyle, by interpolating within the latent space of the GAN. The proposed technique for portrait image animation outperforms previous methods in terms of image quality, identity preservation, and pose transfer while also supporting attribute editing.

CCS Concepts: • Computing methodologies → Image manipulation.

Additional Key Words and Phrases: Portrait Editing, 3D GAN

*Both authors contributed equally.

Project webpage: <https://www.computationalimaging.org/publications/3dganinversion>.
 Authors’ address: Connor Z. Lin; David B. Lindell; Eric R. Chan; Gordon Wetzstein, Stanford University, Stanford, CA, USA, connorzl@stanford.edu, lindell@stanford.edu.

1 INTRODUCTION

Hundreds of millions of images are captured and shared across the world every single day. Many of these are portraits: single-image snapshots of a person’s head. While portrait images are abundant, the appearance and expression of the subject are fixed at the point of capture. The ability to edit a portrait image, or to infer what the subject might look like if captured with a different pose, expression, hairstyle or even physical age, has significant relevance to applications in computer graphics, vision, and virtual reality.

A convincing method for portrait image animation enables image manipulation while preserving the likeness or identity of the subject. When manipulating the pose or posture of an individual, the method should produce multi-view consistent results such that the identity is not altered during the animation.

A number of approaches have been proposed to achieve this goal. For example, methods based on generative adversarial network (GAN) inversion manipulate the latent space of a pre-trained 2D GAN [Choi et al. 2020; Karras et al. 2018, 2021, 2019, 2020; Wang et al. 2018] to adjust expression, pose, lighting, or other attributes [Härkönen et al. 2020; Shen et al. 2020; Tewari et al. 2020a,b]. This approach, however, often results in unpredictable changes to the appearance or identity with changes in viewpoint due to the lack of modeling the 3D structure of the scene. Other methods based on 3D morphable models (3DMMs) [Blanz and Vetter 1999] predict an explicit textured mesh representation of the portrait image subject and can

thus manipulate appearance through animation of the underlying mesh [Deng et al. 2019b; Feng et al. 2021; Kim et al. 2018; Thies et al. 2016]. While successful in editing facial expressions, these approaches are not capable of directly editing hair or semantic attributes. Moreover, these methods cannot usually be directly applied to portrait image animation since they usually require multiple images or a video sequence of a subject in order to synthesize a complete texture map. Finally, another class of methods animates portrait images by warping an input image based on a desired pose and expression and refining the output using a neural network [Ren et al. 2021; Siarohin et al. 2019; Zakharov et al. 2020]. While these methods produce plausible results, they do not achieve the photorealistic quality of GAN-based methods, and generally have view inconsistencies for large deviations from the initial portrait image.

In this work, we address these limitations to achieve multi-view-consistent editing and animation of portrait images. Our approach leverages recently developed 3D-aware GANs, specifically EG3D [Chan et al. 2022], and a 3DMM-based supervision strategy that enables fine-grain control over expression editing. For an input portrait image, we estimate a 3DMM using a pre-trained encoder network [Feng et al. 2021]. Using the 3DMM, we edit the facial expression of the portrait image in a controllable fashion. A main insight is that this edited image can be used to supervise the GAN to produce the same edited expression using a photometric loss. To this end, we perform GAN inversion and fine-tuning to simultaneously embed the 3DMM-synthesized image into the latent space of the GAN and in-paint any masked out regions (e.g., the interior of the mouth). After inversion, the 3D GAN predicts an explicit neural radiance field [Mildenhall et al. 2020], enabling multi-view consistent rendering from arbitrary poses across the edited expressions.

Our approach can similarly be used to drive portrait images for video re-enactment. We use the 3DMM parameters from the video to edit the portrait image expressions and embed each of the resulting frames into the GAN latent space. Likewise, by leveraging the semantic editing capabilities of the GAN, we can control pose while also editing attributes such as age, hairstyle, and gender (see Fig. 1).

In summary, we make the following contributions:

- We propose a method for explicit manipulation of expression and pose of a portrait image by fine-tuning the latent space of a pre-trained 3D GAN with 3DMM-based supervision.
- We extend this approach for video re-enactment from a portrait image and show an improvement in synthesized image quality compared to other methods.
- We demonstrate attribute-edited portrait image animation; that is, we exploit the latent space of the 3D GAN model in order to edit semantic attributes such as age, hairstyle, or gender in addition to explicitly controlling the expression and pose.

Overview of Limitations. Here, we mention a few limitations of our work. First, because our method uses a pre-trained 3D GAN [Chan et al. 2022] trained on a 2D image dataset, the camera pose and head pose are entangled. Thus, manipulating head pose is achieved with a corresponding rotation and translation of the camera, which results in movement of the background and shoulders. While this can potentially be addressed by developing new training techniques

that disentangle head and camera pose, this is out of scope of the current work, as we focus on modeling the appearance of the face and head. Second, we are limited by the characteristics of the FFHQ dataset [Karras et al. 2019] used to train the 3D GAN. As a result, editing eyes is challenging since most GAN-synthesized images have open eyes that are looking directly at the camera.

2 RELATED WORK

In this section, we outline the most relevant and recent work in the area of animating portrait images.

Model-based Portrait Image Editing. Traditionally, related tasks were addressed by fitting 3D morphable models (3DMMs) [Blanz and Vetter 1999; Feng et al. 2021] to a 2D image or video and then editing or re-enacting this model (e.g., [Thies et al. 2019, 2016]). 3DMM-based approaches, however, typically only model the face rather than other details such as hair or teeth. Moreover, these models usually require a texture atlas built from multiple views of the subject to handle disocclusions. A recent survey of these approaches can be found in the state-of-the-art (SoTA) report by Zollhöfer et al. [2018].

Machine Learning-based Portrait Image and Video Editing. Over the last few years, neural synthesis of realistic talking head sequences has become a topic of interest. Autoencoder-type network architectures are often used with videos [Isola et al. 2017; Kim et al. 2018; Lombardi et al. 2018; Wang et al. 2019], few-shot video clips [Ha et al. 2020; Wang et al. 2019; Zakharov et al. 2019], or single-images as input [Averbuch-Elor et al. 2017; Geng et al. 2018; Nagano et al. 2018; Shu et al. 2018; Siarohin et al. 2019; Tewari et al. 2020c, 2021; Tripathy et al. 2019; Wiles et al. 2018; Zakharov et al. 2020]. These approaches can animate a single target image using the motion of a source video. Autoencoder networks, however, make it difficult to attain intuitive parametric control of the animation. Recent approaches aim to overcome this challenge using a keypoint-driven motion flow field [Wang et al. 2021] or a 3DMM [Ren et al. 2021] to drive the portrait image edits. Yet, all these network architectures are oblivious to the underlying 3D structure of the content, making it challenging to produce photorealistic outputs across a range of viewpoints. To address this problem, recent work combines neural radiance field and 3DMMs [Gafni et al. 2021], but this approach requires a video showing the target person from all angles for building an editable representation. In our work, we focus on intuitive and 3DMM-driven editing of a single target portrait image in a multi-view consistent manner.

Portrait Image Editing using GANs. SoTA portrait image editing techniques operate in the latent space of a GAN [Karras et al. 2018, 2021, 2019, 2020; Radford et al. 2016]. The latent space of StyleGAN, in particular, has been studied extensively and enabled image editing using linear and non-linear latent space arithmetic [Abdal et al. 2019, 2020, 2021; Härkönen et al. 2020; Pumarola et al. 2019; Roich et al. 2021; Shen et al. 2020; Tov et al. 2021; Zhu et al. 2020]; local, semantically-aware edits [Collins et al. 2020]; 3DMM-driven control of StyleGAN’s latent space [Tewari et al. 2020a,b]; a combination of traditional face rendering and inverse GANs [Chandran

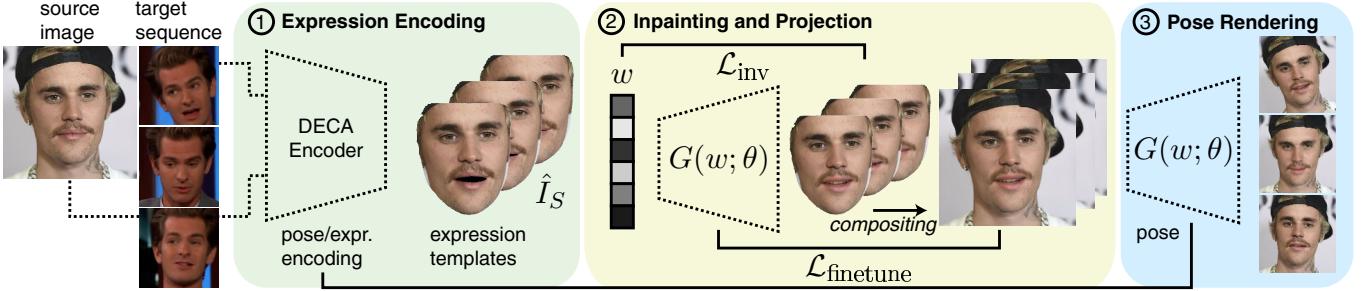


Fig. 2. Overview of the proposed portrait image animation method. Given a source image and target image sequence, our method transfers pose and expression attributes to the source image. (1) We encode the expressions of the target image frames and transfer them to the source image using a 3DMM predicted with a pre-trained model [Feng et al. 2021]. Using the resulting 3DMM, we render expression templates, or images of the source image face with the target expressions. (2) GAN inversion is used to re-render the expression templates, which also in-paints the mouth region if necessary. The in-painted mouth is composited back onto the expression template and image background, and the result is embedded into the GAN latent space using Pivotal Tuning Inversion [Roich et al. 2021]. (3) The final result is rendered by explicitly conditioning the 3D GAN with the poses of the target sequence.

et al. 2021]; and free-viewpoint rendering [Leimkühler and Dretakis 2021] through GAN inversion (see Xia et al. for a survey of GAN inversion techniques [2021]). StyleGAN and related architectures, however, are trained to generate 2D images, which makes it challenging to control head pose, camera perspective, or other 3D attributes in a photorealistic manner. SoFGAN [Chen et al. 2021] recently addressed some of these issues by generating multi-view consistent 3D semantic labels that are textured using a conditional StyleGAN architecture. Yet, this approach is not multi-view consistent as it relies on a conditional 2D GAN to generate the RGB pixel values.

Recent work on 3D GANs show promise in overcoming some of these limitations, but many existing architectures do not achieve the photorealistic image quality of StyleGAN. For example, mesh-based 3D GANs are limited in viewing angle and detail [Liao et al. 2020; Shi et al. 2021; Szabó et al. 2019]; voxel-based 3D GANs are limited in their resolution due to extensive memory requirements [Gadelha et al. 2017; Hao et al. 2021; Henzler et al. 2019; Nguyen-Phuoc et al. 2019, 2020; Wu et al. 2016; Zhu et al. 2018], 3D GANs building on implicit representation networks are computationally inefficient [Chan et al. 2021; Schwarz et al. 2020], and 3D GANs relying on 2D CNN-based image upsampling layers often suffer from view inconsistencies [Gu et al. 2021; Niemeyer and Geiger 2021].

Our work builds on the EG3D architecture [Chan et al. 2022], a recent 3D-aware GAN that is both computationally efficient and offers multi-view consistent image quality almost on par with StyleGAN2. We extend EG3D by exploring its latent space and capabilities of animating a single target image using either intuitive editing of attributes, such as age or gender, or transferring the facial expressions and pose from a target video.

Note that a number of concurrently developed, but still unpublished 3D-aware GANs have recently appeared [Daras et al. 2021; Deng et al. 2021; Gu et al. 2021; Or-El et al. 2021; Sun et al. 2021; Xu et al. 2021; Zhou et al. 2021]. Some of these architectures may also be suitable to animate a single target image, although exploring all of these unpublished works is beyond the scope of this paper.

3 INVERSE 3D GAN FRAMEWORK

In order to animate and edit an input source image, we combine 3DMM-based expression editing with latent space manipulation of a pre-trained 3D GAN. An overview of the method is shown in Fig. 2. Given a target video, we first leverage a 3D face reconstruction model to predict and transfer facial expressions from each target frame to the source image. Then, we perform 3D GAN inversion on the re-expressed source images, projecting them into the latent space of the GAN. This process automatically in-paints any unconstrained regions, such as the interior of the mouth, which are not rendered by the 3DMM. Finally, we use the 3D GAN to re-render the expression-edited source images to match the poses of the target video frames. Our pipeline naturally supports facial attribute editing—the source image is edited using any GAN-based attribute editing method and then re-animated following the same process as above.

3.1 3D GAN

We choose EG3D [Chan et al. 2022] as the pre-trained 3D GAN, as it is the current SOTA for synthesizing photo-realistic, view-consistent images. Given a randomly sampled latent code z , the mapping network maps z to an intermediate latent code w . This intermediate latent code modulates the convolution kernels of a synthesis network to produce tri-plane features. To render the generated image at a desired camera pose, these feature planes are sampled and decoded into a neural radiance field for volumetric rendering. The resulting synthesized images are view-consistent, allowing the 3D GAN to robustly handle occlusion and pose shifts when animating a subject. Please see Chan et al. [2022] for additional details on the 3D GAN.

3.2 Inversion Framework

Given an input source image I_S and a target image I_T , we transfer facial expressions and head pose from I_T to I_S while handling occlusions and preserving the facial identity of I_S .

Expression Transfer. To transfer facial expressions, we employ a pre-trained DECA model [Feng et al. 2021] for 3D face reconstruction. We apply DECA’s encoder to I_S and I_T to obtain the source and target identities ($\beta_S \in \mathbb{R}^{50}, \beta_T \in \mathbb{R}^{50}$), expressions ($\theta_S \in \mathbb{R}^{50}, \theta_T \in \mathbb{R}^{50}$), and poses ($\psi_S \in \mathbb{R}^6, \psi_T \in \mathbb{R}^6$). We then form

an input $(\beta_S, \theta_T, [\psi_{S_{0:3}}, \psi_{T_{3:6}}])$ for DECA’s decoder, consisting of the source’s identity, target’s expression, source’s head pose, and target jaw pose. The decoder predicts a 3DMM that we render using textures sampled from I_S to obtain the expression transferred RGB and face mask output (\hat{I}_S, M_S) . Although we re-use the source’s head pose, swapping in the target’s jaw pose causes the rendered result to have disocclusions in the mouth region when the source’s mouth is closed and the target’s mouth is open. We rely on the 3D GAN to in-paint these regions in a realistic manner during the inversion process. Because DECA does not handle blinking expressions, the proposed pipeline inherits this limitation and does not transfer such eye expressions.

3D GAN Inversion. Given the re-expressed source image and face mask (\hat{I}_S, M_S) from DECA, we draw inspiration from Pivotal Tuning Inversion (PTI) [Roich et al. 2021] and perform 3D GAN inversion in two stages. In the first stage, we optimize for the latent code w that most closely reconstructs the re-expressed face region by minimizing the loss \mathcal{L}_{inv}

$$M_{S_{ij}} = \begin{cases} 1 & \hat{I}_{S_{ij}} \in \text{face} \wedge \hat{I}_{S_{ij}} \notin \text{mouth} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

$$\mathcal{L}_{\text{inv}} = \|\Phi(M_S \odot \hat{I}_S) - \Phi(M_S \odot G(w; \theta))\|_2. \quad (2)$$

Here, Φ is an image feature extractor, such as VGG [Simonyan and Zisserman 2014], and θ are the parameters of the generator, which are frozen during this stage. We limit the optimization to the masked face region to avoid overly constraining the latent space of the generator, as in-the-wild images often contain unseen backgrounds and subjects. During this stage, the generator leverages its prior to realistically hallucinate occlusions near the mouth region.

In the second stage, we freeze the optimized latent code w^* and fine-tune the generator to match the rest of \hat{I}_S . Note that we also freeze the in-painted mouth region, as we find empirically that in-painted regions tend to degenerate unless explicitly constrained throughout the optimization of the generator. We define the loss function used to fine-tune the generator as follows.

$$M_{\text{mouth}_{ij}} = \begin{cases} 1 & \hat{I}_{S_{ij}} \in \text{mouth} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

$$I_{\text{GT}} = (M_{\text{mouth}} \odot \hat{I}_S) + (1 - M_{\text{mouth}}) \odot I_S \quad (4)$$

$$I_{\text{pred}} = G(w^*; \theta), \quad (5)$$

$$\mathcal{L}_{\text{finetune}} = \mathcal{L}_{\text{LPIPS}}(I_{\text{GT}}, I_{\text{pred}}) + \|I_{\text{GT}} - I_{\text{pred}}\|_2, \quad (6)$$

$\mathcal{L}_{\text{LPIPS}}$ is a perceptual loss function [Zhang et al. 2018], M_{mouth} is the mask including all pixels except the mouth region, and θ^* are the fine-tuned generator parameters.

Head Pose Transfer. One major benefit of the 3D GAN is that pose is naturally disentangled and can be easily manipulated while maintaining view consistency. After inversion, we simply re-render I_{pred} using the 3D GAN to match the target head pose, estimated using Deep3DFace [Deng et al. 2019b]. The final output is the edited portrait image with an expression and pose corresponding to the target image.

We show intermediate outputs of this process in Fig. 3. Here, the expression template is the expression-edited source image produced using the 3DMM. Then, after the projection step, we visualize the



Fig. 3. Visualization of intermediate outputs from the image animation pipeline. The expression template is produced using an estimated 3DMM. Then, the face is projected into the latent space of the 3D GAN, which also in-paints the mouth region. After finetuning, the hair and background match the source image and the pose can be manipulated directly by the 3D GAN.

result of rendering w^* using the 3D GAN generator. Note that the face region closely matches the expression template and the mouth region is in-painted. After fine-tuning the generator, regions outside the face, including the hair and background, closely match the original source image. Finally, the pose is adjusted to match the target frames by conditioning the 3D GAN directly with the corresponding pose values.

3.3 Video-based Face Animation

For an input target video V_T consisting of frames $I_{T_0}, I_{T_1}, \dots, I_{T_N}$, we adapt the same expression transfer pipeline to each frame. We smooth the estimated poses over a window of two frames to mitigate camera jittering during head pose transfer. Although the source input is a single image, the 3D GAN naturally facilitates static image animation due to its view consistency. However, subtle flickering artifacts may still arise from the 3D GAN being unable to exactly recover the background and shoulders during inversion. To address this, we apply the re-projected face mask to obtain just the face region F_i for each frame $i > 1$. We then render the first frame using pose i and composite in F_i . Here, we leverage the view consistency of the 3D GAN to ensure that the final result is temporally consistent.

3.4 Facial Attribute Editing

We incorporate attribute editing into the proposed pipeline using an edit-then-fit approach. Specifically, we adapt StyleFlow [Abdal et al. 2021] to work with the pre-trained 3D GAN; given an input w and desired attribute modifications, the method predicts a new w that results in an image with the desired adjustments when rendered using the 3D GAN. This attribute-edited image is used as the input to the rest of the image animation pipeline, described previously.

To adapt StyleFlow for the 3D GAN, we prepare a dataset of 10,000 paired latent codes z and w , sampled randomly from the latent space of the pre-trained 3D GAN. We use the 3D GAN to render the portrait images corresponding to each w , the Microsoft Face API [Microsoft 2020] to generate attribute labels for each image. After assembling this dataset we follow Abdal et al. [2021] to train a continuous normalizing flow [Kobyzev et al. 2020] to provide an invertible mapping between w and z , conditioned on the attributes.

After training, we edit a real input image I by performing PTI to retrieve the w that best reconstructs I and retrieving the image attributes using the Face API. We use the continuous normalizing flow to obtain the original z that maps to the optimized w :

$$z(t_0) = z(t_1) + \int_{t_1}^{t_0} \phi(z(t), t, \alpha; \theta) dt. \quad (7)$$

Here, $z(t_0) = z$, $z(t_1) = w$, α is the vector of attributes, and ϕ is the continuous normalizing flow network. We modify elements of α corresponding to the attributes we wish to edit, forming the updated attribute vector α^* . Finally, we perform a forward flow to produce the latent code $w^* = z(t_1)$ that encodes the edited image I^* with attributes α^* :

$$z(t_1) = z(t_0) + \int_{t_0}^{t_1} \phi(z(t), t, \alpha^*; \theta) dt. \quad (8)$$

To animate the edited source image, we repeat the steps detailed in Sec. 3.3 by synthesizing re-expressed source images using DECA [Feng et al. 2021]. The re-expressed images then serve as input into the 3D GAN inversion pipeline to produce the final outputs.

While StyleFlow or other latent space interpolation techniques [Härkönen et al. 2020; Shen et al. 2020] can be used with 2D GANs to control pose via conditional attributes, the results are limited in pose range and suffer from view inconsistencies and changes to the subject likeness during pose manipulation. Since the 3D GAN predicts a fixed radiance field that can be rendered from any pose, the proposed method is naturally multi-view consistent and preserves the subject likeness better than the 2D approach as we demonstrate in the following section.

4 RESULTS

We present results of the proposed method with comparisons to other baseline methods, including PIRenderer [Ren et al. 2021], a SoTA approach for portrait image animation, and a modified version of the proposed method using a 2D GAN. PIRenderer uses a pre-trained 3DMM encoder to predict the expressions of the target frame and transfers them to the source image using a warping field predicted with a convolutional network. A separate convolutional block then refines the result to produce the final output image. The 2D GAN-based method is a straightforward adaptation of the proposed method, except that pose cannot be directly controlled.

Overall, the 3D GAN provides far better animation quality than the 2D GAN because the pose can be explicitly controlled. While both the 2D GAN and 3D GAN show qualitative and quantitative improvements over PIRenderer in terms of image quality, PIRenderer matches the target expressions more closely, primarily because it is not explicitly constrained by the expressiveness of the face model used for the 2D and 3D GAN. We also show that the pose of the 2D GAN can be manipulated using StyleFlow, but that manipulating the pose using the 3D GAN produces significantly better multi-view consistency and identity preservation across edits.

4.1 Portrait Image Animation

We evaluate the proposed method for portrait image animation and compare against the PIRenderer baseline [Ren et al. 2021] and the proposed approach using a pre-trained 2D GAN [Karras et al. 2020].

Implementation Details. The 3D GAN is adapted directly from Chan et al. [Chan et al. 2022] with pre-training on the Flicker-Faces-HQ (FFHQ) dataset [Karras et al. 2019] and an output resolution of 512×512 pixels. We adapt DECA to perform the expression encoding with some modifications to the original codebase [Feng et al. 2021] to render the predicted 3DMM face using a high-resolution texture map extracted from the input source image. After rendering out the face masks using DECA, we optimize w for the first frame in the image sequence for 2,000 steps using the Adam optimizer [Kingma and Ba 2014]. For subsequent frames, we initialize w with the value from the previous frame and run the optimization for an additional 100 steps. We fine-tune the generator for 10,000 steps using a batch size of 1. The whole procedure takes roughly two hours for a sequence of 100 target images running on an NVIDIA A6000 GPU. To improve temporal consistency in the final result, we re-render the first frame of the sequence with the pose of each subsequent frame. For all other frames, we extract the rendered face and composite it back onto the first-frame image with the same pose. This procedure removes any jitter or variation in the background pixels between frames, and the face extraction can be implemented automatically by re-projecting the 3DMM face mask onto the target pose using the depth maps predicted by the 3D GAN. We will make all code publicly available prior to publication.

Qualitative Results. We show qualitative comparisons of the proposed method, PIRenderer, and the approach using the 2D GAN in Fig. 4 with additional results in the supplemental. PIRenderer shows good control over the expressions, including the mouth and eyes of the source image. However, the output images tend to be lower fidelity than results from the 2D or 3D GAN, with blurrier details and patchy artifacts where the network attempts to in-paint missing details. Moreover, since there is no explicit 3D view consistency, the identity of the source image may shift as the pose changes. Qualitatively, the 2D and 3D GAN results are similar, but since there is no native mechanism for manipulating the 2D GAN pose, only the expression is modulated. Since the 3DMM framework cannot be used to synthesize eye blinks and both the 2D GAN and 3D GAN are trained on FFHQ [Karras et al. 2019], a dataset in which few subjects are blinking, the proposed framework cannot adjust the eye position or blink state. Still, the 3D GAN achieves a good match to the target pose with similar expressions while reproducing fine details of the input image. Importantly, the 3D GAN also preserves the identity of the input image during pose adjustments. Note that the pretrained and 2D and 3D GANs require slightly different image crops, and so we crop the images according to their respective implementations (the 2D GAN crop is somewhat more centered on the entire head, see Fig. 4).

Quantitative Results. We observe similar trends in the quantitative results, reported in Table 1. These results are calculated over a dataset of 500 images from CelebA-HQ [Karras et al. 2018] and 500 target video sequences sampled from VoxCeleb [Chung et al. 2018;

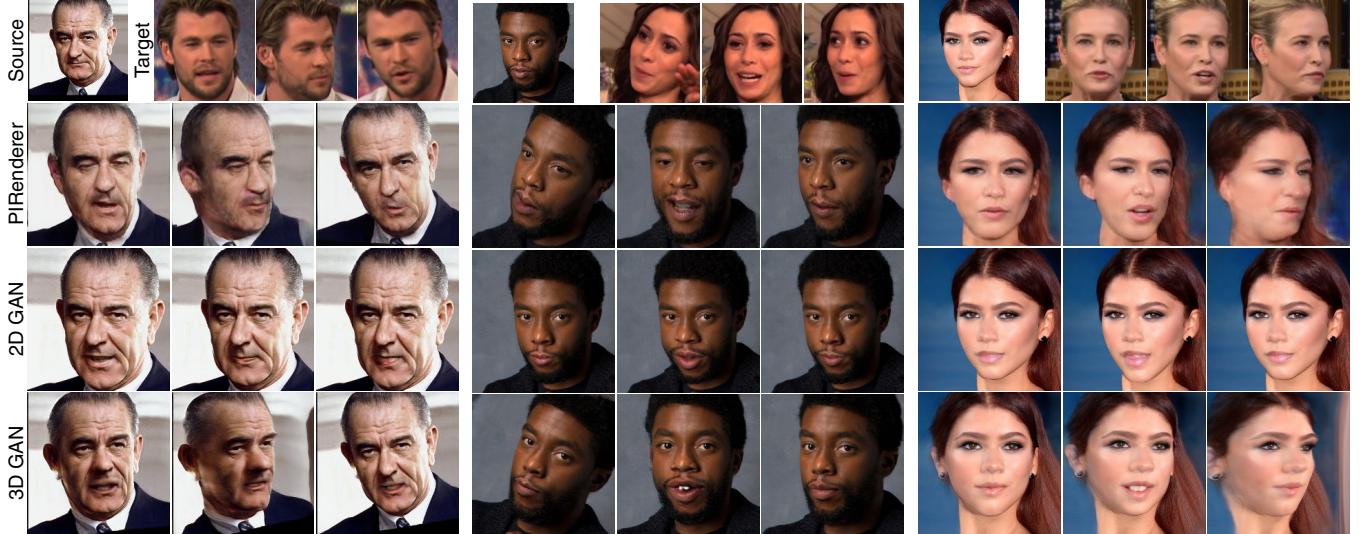


Fig. 4. Portrait image animation results. We compare the proposed approach with PIRenderer [Ren et al. 2021] and the proposed approach using a pre-trained 2D GAN [Karras et al. 2020] instead of the 3D GAN. Since we cannot directly control the pose with the 2D GAN, we only edit the expressions. The proposed 3D GAN-based approach achieves better view consistency with higher output image quality compared to PIRenderer.

	FID↓	ID ↑	APD↓	AED↓
PIRenderer (w/o eyes, w/o pose)	53.916	-	0.250	0.437
PIRenderer (w/o pose)	53.959	-	0.247	0.386
PIRenderer (w/o eyes)	63.844	0.694	0.039	0.424
PIRenderer	64.379	0.700	0.040	0.373
2D GAN (w/o pose)	17.812	-	0.246	0.434
3D GAN (w/o pose)	16.504	-	0.246	0.433
3D GAN	31.176	0.733	0.030	0.433

Table 1. Quantitative evaluation of portrait image animation. We compare our approach using the pre-trained 3D GAN of Chan et al. [2022] to our approach using a 2D GAN [Karras et al. 2020] and PIRenderer [Ren et al. 2021]. The 3D GAN achieves the best performance in terms of Fréchet Inception Distance (FID), identity consistency (ID), avg. pose distance (APD), while PIRenderer achieves the best avg. expression distance (AED).

PIRenderer	2D GAN				3D GAN	
	w/o eyes, w/o pose	w/o pose	w/o eyes	full	w/o pose	full
0.652	0.622	0.476	0.451	0.891	0.896	0.590

Table 2. Evaluation of identity preservation compared to the source image. The provided ArcFace similarity score [Deng et al. 2019a] has values between -1 and 1 (greater value means more similar) and is calculated by comparing the rendered images to the input source image. Scores improve when removing pose or eye modeling, but the 3D GAN produces the best results compared to relevant variants of PIRenderer or the 2D GAN.

Nagrani et al. 2019, 2017]. We evaluate using the Fréchet Inception Distance (FID), average pose distance (APD), average expression distance (AED), and an identity consistency metric (ID).

FID is computed between the original 500 source images and 500 samples of the edited source images, averaged over the 100 possible realizations of these sets of images, sampled without replacement. For each video clip, the APD and AED are computed between the

modified source images and the target images. The APD is given as root mean squared error between the estimated pose encodings from a pre-trained Deep3DFace model [Deng et al. 2019b] for the modified source image and the target image sequence. The AED is defined similarly, except the metric is calculated between the expression encodings predicted by the pre-trained model. Finally, we also compute an identity consistency metric using a pre-trained Arcface model [Deng et al. 2019a]. For this metric, we randomly sample 2,000 poses from the CelebA-HQ dataset and randomly apply two poses to two different frames from each video clip for a total of 1000 image pairs. We calculate the cosine distance of the predicted embedding for each pair and report the average result. To evaluate how well the identity is preserved compared to the source embedding, we randomly sample two rendered frames corresponding to each of the 500 source images and compute the identity consistency metric between these samples and their source images. Results are shown in Table 2

The proposed method using the 3D GAN achieves the best accuracy in terms of FID score, average pose distance, and identity consistency (see Table 1). We hypothesize that the limited ability of our model to capture eye motion (e.g., blinking) results in a penalty to the expression distance metric. A similar drop in performance is observed for a variant of PIRenderer with fixed eye expressions (see “w/o eyes” in Table 1), which we achieve by fixing the eye vertices of the 3DMM model to those of the first frame in the target image sequence. Finally, since the 2D GAN does not explicitly model pose, we observe that its performance is similar to a variant of the 3D GAN with fixed poses or PIRenderer where poses and eye expressions are fixed. Note that methods with fixed pose (“w/o pose”) have improved FID scores because they are not animated away from the initial poses of the ground truth image dataset. Similarly, variants

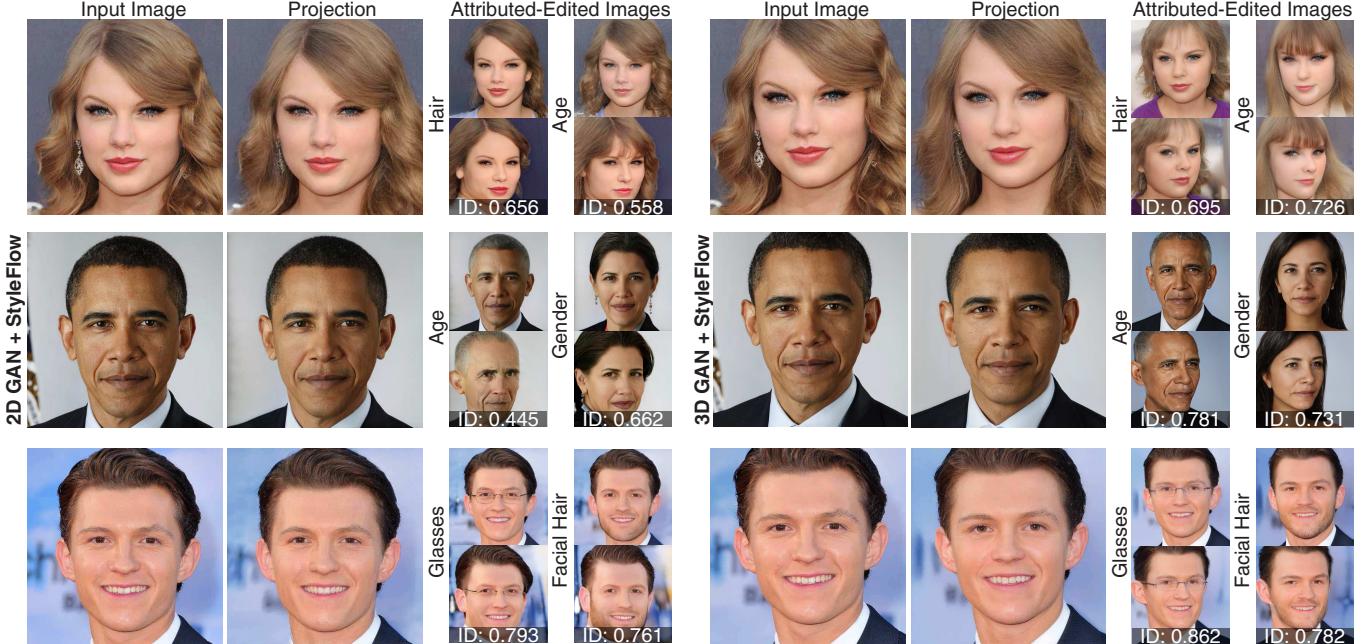


Fig. 5. Attribute editing results. We compare attribute editing and portrait image manipulation using the 2D GAN [Karras et al. 2020] and the 3D GAN [Chan et al. 2022]. We perform attribute edits using StyleFlow [Abdal et al. 2021] and adapt this method for the 3D GAN. Qualitatively both GANs achieve similarly expressive edits, but the 3D GAN demonstrates greater multi-view consistency when adjusting the pose after editing, as shown by the 3D GAN’s higher ID similarity scores [Deng et al. 2019a]. The two GANs expect slightly different input crops and pose is controlled for the 2D GAN via StyleFlow editing.

	Pose Only	+ Age	+ Facial Hair	+ Gender	+ Glasses	+ Hair
2D GAN	0.652	0.637	0.638	0.652	0.653	0.686
3D GAN	0.790	0.797	0.790	0.793	0.794	0.801

Table 3. Quantitative evaluation of multi-view consistency and face identity preservation via ArcFace cosine similarity scores [Deng et al. 2019a] – a score of -1 and 1 indicate maximum dissimilarity or similarity, respectively.

of the methods with fixed pose or eyes have improved identity consistency for the same reason (Table 2), though the 3D GAN shows the best consistency among the relevant baseline variants.

4.2 Attribute Editing

Implementation Details. We use a pre-trained StyleGAN2 model for the 2D GAN baseline and a pre-trained EG3D model for our 3D GAN. We re-train a StyleFlow model for each method following the original paper. To generate each training dataset, we randomly sample 10,000 w latent vectors with a truncation factor of 0.7 to reduce the number of low quality outlier faces. Both StyleFlow models are trained for 20 epochs using a batch size of 5. To perform edits for an input image, we first perform Pivotal Tuning Inversion [Roich et al. 2021] with 500 w optimization steps and only 100 generator fine-tuning steps with locality regularization to minimize latent space distribution shift.

Qualitative Results. Fig. 5 compares attribute editing results between the 2D GAN and 3D GAN using StyleFlow to control hair, age, glasses, gender, and facial hair. Qualitatively, both the 2D GAN and the 3D GAN can reconstruct the input image faithfully and

exhibit similarly reasonable editing results for frontal facing subjects. In both cases, we observe some shift in the identity, e.g., when modifying hairstyle which seems to be a limitation of the StyleFlow editing directions and perhaps the expressiveness of the GAN latent spaces. However, the 3D GAN is significantly more view-consistent, as the 2D GAN undergoes a large degree of identity shift when the pitch and yaw attributes are shifted.

Quantitative Results. We evaluate multi-view consistency and face identity preservation by measuring ArcFace cosine similarity. As shown by Fig. 5, editing with the 3D GAN produces higher similarity scores than with the 2D GAN. Table 3 demonstrates that this trend also holds for generated subjects. We randomly sample 128 subjects from each GAN’s latent space, apply an attribute edit to each subject, and render out two views for each subject. These poses are uniformly sampled from yaw angles between $[-30^\circ, 30^\circ]$ and pitch angles between $[-20^\circ, 20^\circ]$.

5 DISCUSSION

We have presented a method for controllable portrait image animation based on 3D GAN inversion. The proposed approach combines explicit control over expressions afforded by 3DMMs with the abilities of emerging 3D GANs to edit semantic information, in-paint missing details, and explicitly control the pose of rendered portrait images. Overall, leveraging a 3D GAN for portrait image animation significantly improves image quality over non-generative approaches and multi-view consistency over 2D GAN methods.

Still, there are a number of limitations to the proposed method. Firstly, the method is constrained by the underlying face model. A direct consequence is that the current implementation does not support rendering eye blinks or changes to the eye position. We also acknowledge limitations of the 3D GAN used in the method; the FFHQ dataset used to train the 3D GAN is not calibrated to represent a balanced cross-section of human appearances. As a result, portrait animation and attribute editing performance is somewhat dependent on the input image appearance. These limitations could be alleviated by using PCA-based editing techniques [Härkönen et al. 2020] to synthesize blinking textures, more sophisticated face tracking models, or improving the GAN training procedure to disentangle the subject expression, allowing explicit control similar to what is currently possible for the pose.

Overall, this work takes important steps towards improving the quality of portrait image animation and the utility of pre-trained GANs for controllable image manipulation. We envision that this work could have relevance to applications across graphics and interactive techniques, including for social media, virtual telecommunications, automated talking head synthesis, and realistic re-dubbing for television and film.

Societal Impact. The proposed method enables the synthesis and animation of photorealistic images; recent advances in this area have enabled new and incredible applications, but also have potential for misuse. We condemn the use of this work for ill-intentioned purposes, not limited to misinformation or impersonation, and support the research of methods to thwart such efforts. We refer the reader to Tewari et al. [2020c] for an extended discussion of methods for detecting and analyzing deepfake videos.

ACKNOWLEDGMENTS

This project was supported in part by a PECASE by the ARO, NSF award 1839974, Stanford HAI, and a Samsung GRO.

REFERENCES

- Rameen Abdal, Yipeng Qin, and Peter Wonka. 2019. Image2StyleGAN: How to Embed Images into the StyleGAN Latent Space?. In *IEEE International Conference on Computer Vision (ICCV)*.
- Rameen Abdal, Yipeng Qin, and Peter Wonka. 2020. Image2StyleGAN++: How to Edit the Embedded Images?. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. 2021. StyleFlow: Attribute-conditioned Exploration of StyleGAN-Generated Images using Conditional Continuous Normalizing Flows. *ACM Transactions on Graphics (ToG)* 40, 3 (2021).
- Hadar Averbuch-Elor, Daniel Cohen-Or, Johannes Kopf, and Michael F. Cohen. 2017. Bringing Portraits to Life. *ACM Transactions on Graphics (ToG)* 36, 6 (2017).
- Volkert Blanz and Thomas Vetter. 1999. A Morphable Model for the Synthesis of 3D Faces. In *Proceedings of SIGGRAPH*.
- 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. 2022. Efficient Geometry-aware 3D Generative Adversarial Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. 2021. pi-GAN: Periodic Implicit Generative Adversarial Networks for 3D-Aware Image Synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Prashanth Chandran, Sebastian Winberg, Gaspard Zoss, Jérémie Rivière, Markus Gross, Paulo Gotardo, and Derek Bradley. 2021. Rendering with Style: Combining Traditional and Neural Approaches for High-Quality Face Rendering. *ACM Transactions on Graphics (ToG)* 40, 6 (2021).
- Anpei Chen, Ruiyang Liu, Ling Xie, Zhang Chen, Hao Su, and Yu Jingyi. 2021. SofGAN: A Portrait Image Generator with Dynamic Styling. *ACM Trans. Graph.* 41, 1 (2021).
- Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. 2020. StarGAN v2: Diverse Image Synthesis for Multiple Domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- J. S. Chung, A. Nagrani, and A. Zisserman. 2018. VoxCeleb2: Deep Speaker Recognition. In *Interspeech*.
- Edo Collins, Raja Bala, Bob Price, and Sabine Süstrunk. 2020. Editing in Style: Uncovering the Local Semantics of GANs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Giannis Daras, Wen-Sheng Chu, Abhishek Kumar, Dmitry Lagun, and Alexandros G. Dimakis. 2021. Solving Inverse Problems with NerfGANs. *arXiv:2112.09061 [cs.CV]*
- Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. 2019a. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *CVPR*.
- Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. 2021. GRAM: Generative Radiance Manifolds for 3D-Aware Image Generation. In *arXiv*.
- Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. 2019b. Accurate 3D Face Reconstruction with Weakly-Supervised Learning: From Single Image to Image Set. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. 2021. Learning an Animatable Detailed 3D Face Model from In-The-Wild Images. *ACM Transactions on Graphics (SIGGRAPH)* 40, 8.
- Matheus Gadelha, Subhransu Maji, and Rui Wang. 2017. 3D Shape Induction from 2D Views of Multiple Objects. In *International Conference on 3D Vision*.
- Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2021. Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Jiahao Geng, Tianjia Shao, Youyi Zheng, Yanlin Weng, and Kun Zhou. 2018. Warp-Guided GANs for Single-Photo Facial Animation. *ACM Transactions on Graphics (ToG)* 37, 6 (2018).
- Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. 2021. StyleNeRF: A Style-based 3D-Aware Generator for High-resolution Image Synthesis. *arXiv preprint arXiv:2110.08985* (2021).
- Sungjoo Ha, Martin Kersner, Beomsu Kim, Seokjun Seo, and Dongyoung Kim. 2020. MarioNETte: Few-shot Face Reenactment Preserving Identity of Unseen Targets. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Zekun Hao, Arun Mallya, Serge Belongie, and Ming-Yu Liu. 2021. GANcraft: Unsupervised 3D Neural Rendering of Minecraft Worlds. In *IEEE International Conference on Computer Vision (ICCV)*.
- Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. 2020. GANSpace: Discovering Interpretable GAN Controls. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Philipp Henzler, Niloy J Mitra, and Tobias Ritschel. 2019. Escaping Plato’s Cave: 3D Shape From Adversarial Rendering. In *IEEE International Conference on Computer Vision (ICCV)*.
- Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. *CVPR* (2017).
- Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *International Conference on Learning Representations (ICLR)*.
- Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2021. Alias-free Generative Adversarial Networks. *Advances in Neural Information Processing Systems* 34 (2021).
- Tero Karras, Samuli Laine, and Timo Aila. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Hyeyeonwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Nießner, Patrick Pérez, Christian Richardt, Michael Zollöfer, and Christian Theobalt. 2018. Deep Video Portraits. *ACM Transactions on Graphics (TOG)* 37, 4 (2018).
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations (ICLR)*.
- Ivan Kobyzev, Simon Prince, and Marcus Brubaker. 2020. Normalizing Flows: An Introduction and Review of Current Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2020).
- Thomas Leimkühler and George Drettakis. 2021. FreeStyleGAN: Free-view Editable Portrait Rendering with the Camera Manifold. 40, 6 (2021).
- Yiyi Liao, Katja Schwarz, Lars Mescheder, and Andreas Geiger. 2020. Towards Unsupervised Learning of Generative Models for 3D Controllable Image Synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. 2018. Deep Appearance Models for Face Rendering. *ACM Trans. Graph.* 37, 4, Article 68 (July 2018), 13 pages.
- Microsoft. 2020. Azure Face. <https://azure.microsoft.com/en-in/services/cognitive-services/face/>

- Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *European Conference on Computer Vision (ECCV)*.
- Koki Nagano, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, and Hao Li. 2018. PaGAN: Real-Time Avatars Using Dynamic Textures. *ACM Transactions on Graphics (ToG)* 37, 6 (2018).
- Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. 2019. VoxCeleb: Large-scale Speaker Verification in the Wild. *Computer Science and Language* (2019).
- A. Nagrani, J. S. Chung, and A. Zisserman. 2017. VoxCeleb: A Large-scale Speaker Identification Dataset. In *Interspeech*.
- Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. 2019. HoloGAN: Unsupervised Learning of 3D Representations from Natural Images. In *IEEE International Conference on Computer Vision (ICCV)*.
- Thu Nguyen-Phuoc, Christian Richardt, Long Mai, Yong-Liang Yang, and Niloy Mitra. 2020. BlockGAN: Learning 3D Object-aware Scene Representations from Unlabelled Images. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Michael Niemeyer and Andreas Geiger. 2021. GIRAFFE: Representing Scenes as Compositional Generative Neural Feature Fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. 2021. StyleSDF: High-Resolution 3D-Consistent Image and Geometry Generation. *arXiv preprint arXiv:2112.11427* (2021).
- A. Pumarola, A. Agudo, A.M. Martinez, A. Sanfelix, and F. Moreno-Noguer. 2019. GANimation: One-Shot Anatomically Consistent Facial Animation. (2019).
- Alec Radford, Luke Metz, and Soumith Chintala. 2016. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In *International Conference on Learning Representations (ICLR)*.
- Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. 2021. PIRenderer: Controllable Portrait Image Generation via Semantic Neural Rendering. In *IEEE International Conference on Computer Vision (ICCV)*.
- Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. 2021. Pivotal Tuning for Latent-based Editing of Real Images. *arXiv preprint arXiv:2106.05744* (2021).
- Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. 2020. GRAF: Generative Radiance Fields for 3D-Aware Image Synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Yujun Shen, Jinjin Gu, Xiaou Tang, and Bolei Zhou. 2020. Interpreting the Latent Space of GANs for Semantic Face Editing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Yichun Shi, Divyansh Aggarwal, and Anil K Jain. 2021. Lifting 2D StyleGAN for 3D-Aware Face Generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Zhixin Shu, Mihir Sahasrabudhe, Riza Alp Guler, Dimitris Samaras, Nikos Paragios, and Iasonas Kokkinos. 2018. Deforming Autoencoders: Unsupervised Disentangling of Shape and Appearance. In *European Conference on Computer Vision*.
- Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First Order Motion Model for Image Animation. In *Conference on Neural Information Processing Systems (NeurIPS)*.
- Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-scale Image Recognition. *arXiv preprint arXiv:1409.1556* (2014).
- Jingxiang Sun, Xuan Wang, Yong Zhang, Xiaoyu Li, Qi Zhang, Yebin Liu, and Jue Wang. 2021. FENeRF: Face Editing in Neural Radiance Fields. *arXiv:2111.15490 [cs.CV]*
- Attila Szabó, Givi Meishvili, and Paolo Favaro. 2019. Unsupervised Generative 3D Shape Learning from Natural Images. *arXiv preprint arXiv:1910.00287* (2019).
- Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zöllhofer, and Christian Theobalt. 2020a. StyleRig: Rigging StyleGAN for 3D Control over Portrait Images, CVPR 2020. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Ayush Tewari, Mohamed Elgharib, Mallikarjun BR, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zöllhofer, and Christian Theobalt. 2020b. PIE: Portrait Image Embedding for Semantic Control. *ACM Transactions on Graphics (Proceedings SIGGRAPH Asia)* 39, 6. <https://doi.org/10.1145/3414685.3417803>
- Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason Saragih, Matthias Nießner, et al. 2020c. State of the Art on Neural Rendering. *Eurographics Association* (2020).
- Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, Yifan Wang, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, Tomas Simon, Christian Theobalt, Matthias Niessner, Jonathan T. Barron, Gordon Wetzstein, Michael Zollhoefer, and Vladislav Golyanik. 2021. Advances in Neural Rendering. *arXiv preprint arXiv:2111.05849* (2021).
- Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2019. Deferred Neural Rendering: Image Synthesis using Neural Textures. *ACM Transactions on Graphics (ToG)* 38, 4 (2019).
- Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2Face: Real-time Face Capture and Reenactment of RGB Videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. 2021. Designing an Encoder for StyleGAN Image Manipulation. *ACM Transactions on Graphics (ToG)* 40, 4 (2021).
- Soumya Tripathy, Juho Kannala, and Esa Rahtu. 2019. ICface: Interpretable and Controllable Face Reenactment Using GANs. *arXiv preprint arXiv:1904.01909* (2019).
- Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. 2019. Few-shot Video-to-Video Synthesis. In *Conference on Neural Information Processing Systems (NeurIPS)*.
- Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. 2021. One-Shot Free-View Neural Talking-Head Synthesis for Video Conferencing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- O. Wiles, A.S. Koepke, and A. Zisserman. 2018. X2Face: A Network for Controlling Face Generation by using Images, Audio, and Pose Codes. In *European Conference on Computer Vision*.
- Jiajun Wu, Chengkai Zhang, Tianfan Xue, William T. Freeman, and Joshua B. Tenenbaum. 2016. Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. 2021. GAN inversion: A survey. *arXiv preprint arXiv:2101.05278* (2021).
- Yinghao Xu, Sida Peng, Ceyuan Yang, Yujun Shen, and Bolei Zhou. 2021. 3D-aware Image Synthesis via Learning Structural and Textural Representations. (2021).
- Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. 2020. Fast Bi-layer Neural Synthesis of One-Shot Realistic Head Avatars. In *European Conference on Computer Vision (ECCV)*.
- Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. 2019. Few-Shot Adversarial Learning of Realistic Neural Talking Head Models. *arXiv:1905.08233*
- Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. 2021. CIPS-3D: A 3D-Aware Generator of GANs Based on Conditionally-Independent Pixel Synthesis. *arXiv preprint arXiv:2110.09788* (2021).
- Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. 2020. In-domain GAN Inversion for Real Image Editing. In *European Conference on Computer Vision (ECCV)*.
- Jun-Yan Zhu, Zhoutong Zhang, Chengkai Zhang, Jiajun Wu, Antonio Torralba, Joshua B. Tenenbaum, and William T. Freeman. 2018. Visual Object Networks: Image Generation with Disentangled 3D Representations. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. 2018. State of the Art on Monocular 3D Face Reconstruction, Tracking, and Applications. In *Computer Graphics Forum*, Vol. 37.

3D GAN Inversion for Controllable Portrait Image Animation

Supplemental Material

CONNOR Z. LIN*, DAVID B. LINDELL*, ERIC R. CHAN, and GORDON WETZSTEIN, Stanford University, USA

S1 SUPPLEMENTARY PORTRAIT IMAGE ANIMATION RESULTS

Here we provide additional results and details on the portrait image animation examples presented in the main paper.

Quantitative Results. In Table 1, we further breakdown the average pose distance and average expression distance metrics into separate evaluations using the pre-trained models of DECA [?] and Deep3DFace [?]. We observe the same trends across both estimators, as PIRenderer achieves more accurate expressions but performs similarly to the 3D GAN when eye expressions are fixed.

Qualitative Results. Figure S1 demonstrates an additional result of editing an input source image and animating the result to follow the expressions of a target video. We also show more portrait image animation comparisons with PIRenderer and the 2D GAN baselines in Figure S2. Finally, Figure S3 shows a best-effort comparison with Face-Vid2Vid [?], as only a limited online demo is available for controlling eye and head pose¹.

S2 SUPPLEMENTARY ATTRIBUTE EDITING RESULTS

Qualitative Results. In Figure S4, we illustrate more attribute editing results using StyleFlow [?] and the 3D GAN. Similar to the 2D GAN results shown in the original StyleFlow paper, we observe that the latent space of the 3D GAN facilitates realistic attribute editing. Facial hair edits for female subjects and glasses edits incur the most identity shift due to the sparsity of such examples in FFHQ, the 3D GAN’s training dataset.

*Both authors contributed equally.

¹<http://imaginaires.cc/vid2vid-cameo/>

	FID↓	ID ↑	Avg. Pose Dist.↓ DECA	Avg. Pose Dist.↓ D3DF	Avg. Expression Dist.↓ DECA	Avg. Expression Dist.↓ D3DF
PIRenderer (w/o eyes, w/o pose)	53.916	-	0.183	0.250	0.361	0.437
PIRenderer (w/o pose)	53.959	-	0.179	0.247	0.299	0.386
PIRenderer (w/o eyes)	63.844	0.694	0.043	0.039	0.354	0.424
PIRenderer	64.379	0.700	0.040	0.040	0.282	0.373
2D GAN (w/o pose)	17.812	-	0.179	0.246	0.369	0.434
3D GAN (w/o pose)	16.504	-	0.180	0.246	0.369	0.433
3D GAN	31.176	0.733	0.039	0.030	0.362	0.433

Table 1. Quantitative evaluation of portrait image animation. We compare our approach using the pre-trained 3D GAN of Chan et al. [?] to our approach using a 2D GAN [?] and PIRenderer [?]. The average pose and expression distances are calculated using the estimated pose and expression encodings from the pre-trained models of DECA [?] and Deep3DFace [?], and identity consistency (ID) is calculated using a pre-trained Arcface model [?]. The proposed method using the 3D GAN achieves the best accuracy in terms of FID score, average pose distance, and identity consistency. We hypothesize that the limited ability of our model to capture eye motion (e.g., blinking) results in a penalty to the expression distance metric. A similar drop in performance is observed for a variant of PIRenderer with fixed eye expressions (w/o eyes). Finally, since the 2D GAN does not model pose, we observe that its performance is similar to a variant of the 3D GAN with fixed poses or PIRenderer where poses and eye expressions are fixed. Note that methods with fixed pose (w/o pose) have improved FID scores because they are not animated away from the initial poses of the ground truth image dataset.



Fig. S1. Portrait image animation and attribute editing using the proposed technique. Given a source portrait image and a target expression (e.g., specified with a target image), our method transfers the expression and pose to the input source image. We achieve multi-view consistent edits of pose by embedding the expression-edited portrait image into the latent space of a 3D GAN. By interpolating within the latent space of the GAN, we can also apply our method to animate attribute-edited images, allowing adjustments to age, hair, gender or appearance in addition to expression and pose.



Fig. S2. Portrait image animation results. We compare the proposed approach with PIRenderer [?] and our method using a pre-trained 2D GAN [?] instead of the 3D GAN. Since we cannot directly control the pose with this method, we only edit the expressions. Compared to the baselines, the proposed approach achieves better view consistency with higher output image quality compared to PIRenderer.



Fig. S3. Portrait image animation results. We compare the proposed approach with PIRenderer [?], Face-Vid2Vid [?], and our method using a pre-trained 2D GAN [?] instead of the 3D GAN. Only the online demo is available for Face-Vid2Vid, so we can only control the head and eye poses. We cannot directly control the pose with the 2D GAN, so we only edit the expressions. Compared to the baselines, the proposed approach achieves better view consistency with higher output image quality compared to PIRenderer, and higher head pose and mouth expression accuracy compared to Face-Vid2Vid.

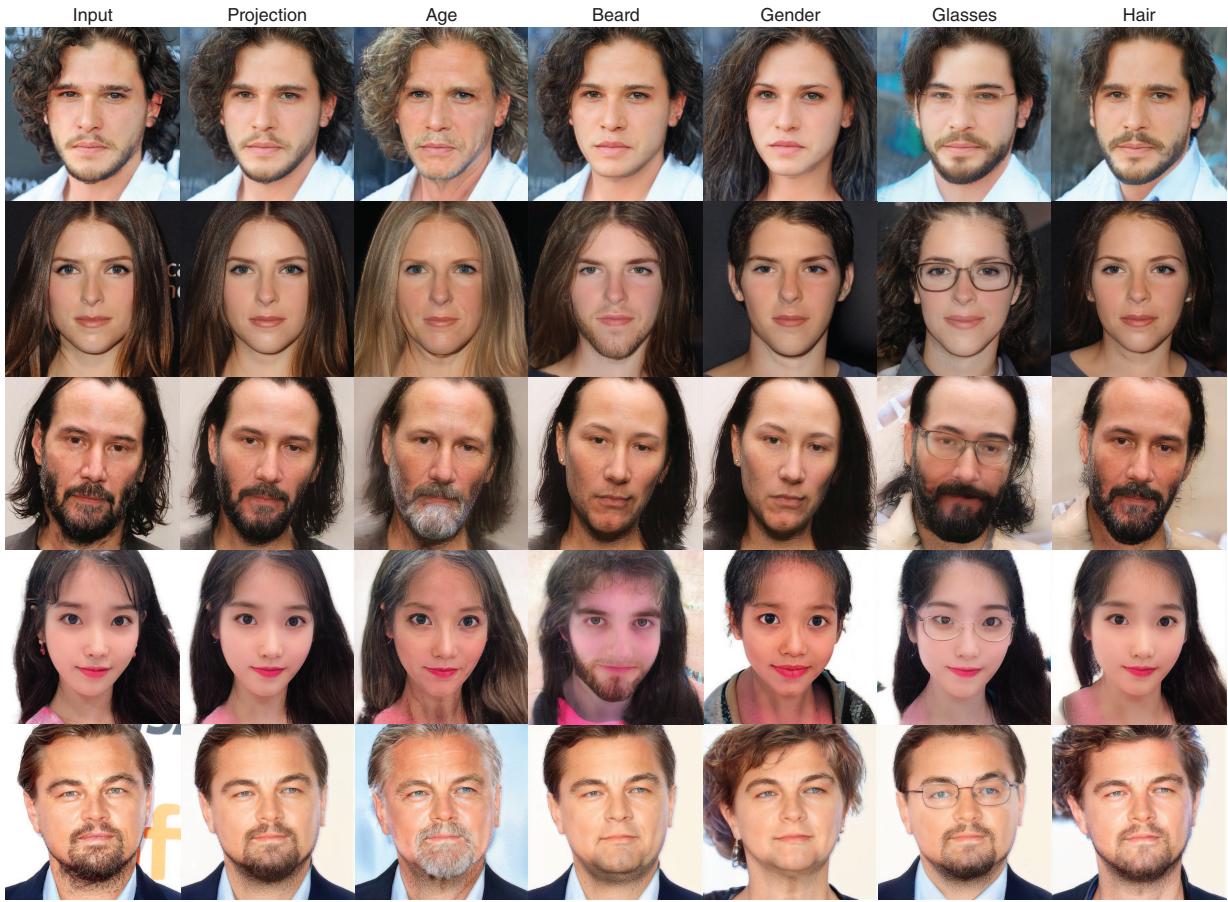


Fig. S4. Portrait image attribute editing results. We train a StyleFlow model [?] using the 3D GAN latent space, then perform Pivotal Tuning Inversion [?] to project real images into the latent space. Finally, we modify the conditional attribute inputs into the StyleFlow model to obtain latents capturing the desired edits. We notice that the most difficult edits are facial hair for female subjects and glasses, as both tend to incur the most identity shift – the former involving gender and the latter involving hair.