GIF: Generative Interpretable Faces

Partha Ghosh[1]   Pravir Singh Gupta*[2]   Roy Uziel*[3]

Anurag Ranjan[1]   Michael J. Black[1]   Timo Bolkart[1]

[1] Max Planck Institute for Intelligent Systems, Tübingen, Germany

[2] Texas A&M University, College Station, USA   [3] Ben Gurion University, Be'er Sheva, Israel

{pghosh,aranjan,black,tbolkart}@tue.mpg.de   pravir@tamu.edu   uzielr@post.bgu.ac.il
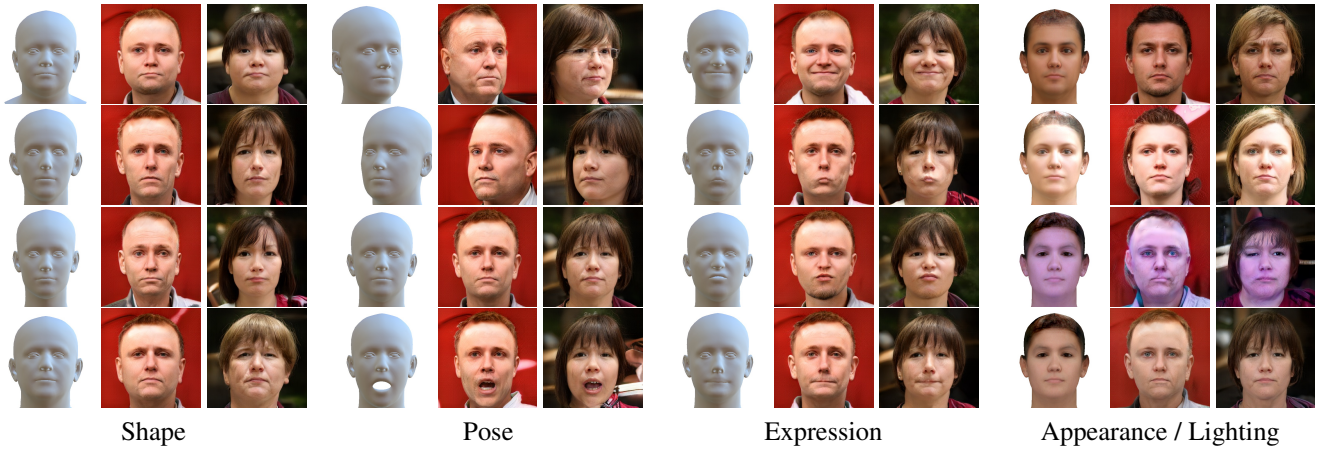
Figure 1: Face images generated by controlling FLAME [38] parameters, appearance parameters, and lighting parameters. For shape and expression, two principal components are visualized at ± 3 standard deviations. The pose variations are visualized at $\pm\pi/8$ (head pose) and at $0, \pi/12$ (jaw pose). For shape, pose, and expression, the two columns are generated for two randomly chosen sets of appearance, lighting, and style parameters. For the appearance and lighting variations (right), the top two rows visualize the first principal components of the appearance space at ± 3 standard deviations, the bottom two rows visualize the first principal component of the lighting parameters at ± 2 standard deviations. The two columns are generated for two randomly chosen style parameters.

Abstract

Photo-realistic visualization and animation of expressive human faces have been a long standing challenge. 3D face modeling methods provide parametric control but generates unrealistic images, on the other hand, generative 2D models like GANs (Generative Adversarial Networks) output photo-realistic face images, but lack explicit control. Recent methods gain partial control, either by attempting to disentangle different factors in an unsupervised manner, or by adding control post hoc to a pre-trained model. Unconditional GANs, however, may entangle factors that are hard to undo later. We condition our generative model on pre-defined control parameters to encourage disentanglement in the generation process. Specifically,

we condition StyleGAN2 on FLAME, a generative 3D face model. While conditioning on FLAME parameters yields unsatisfactory results, we find that conditioning on rendered FLAME geometry and photometric details works well. This gives us a generative 2D face model named GIF (Generative Interpretable Faces) that offers FLAME's parametric control. Here, interpretable refers to the semantic meaning of different parameters. Given FLAME parameters for shape, pose, expressions, parameters for appearance, lighting, and an additional style vector, GIF outputs photo-realistic face images. We perform an AMT based perceptual study to quantitatively and qualitatively evaluate how well GIF follows its conditioning. The code, data, and trained model are publicly available for research purposes at http://gif.is.tue.mpg.de.

---

*Equal contribution.

1

1. Introduction

The ability to generate a person's face has several uses in computer graphics and computer vision that include constructing a personalized avatar for multimedia applications, face recognition and face analysis. To be widely useful, a generative model must offer control over the generative factors such as expression, pose, shape, lighting, skin tone, etc. Early work focuses on learning a low dimensional representation of human faces using Principal Component Analysis (PCA) spaces [13, 14, 15, 16, 63] or higher-order tensor generalizations [65]. Although they provide some semantic control, these methods use linear transformations in the pixel-domain to model facial variation, resulting in blurry images. Further, effects of rotations are not well parameterized by linear transformations in 2D, resulting in poor image quality.

To overcome this, Blanz and Vetter [7] introduced a statistical 3D morphable model, of facial shape and appearance. Such a statistical face model (e.g. [12, 38, 47]) is used to manipulate shape, expression, or pose of the facial 3D geometry. This is then combined with a texture map (e.g. [53]) and rendered to an image. Rendering of a 3D face model lacks photo-realism due to the difficulty in modeling hair, eyes, and the mouth cavity (i.e., teeth or tongue), along with the absence of facial details like wrinkles in the geometry of the facial model. Further difficulties arise from subsurface scattering of facial material. These factors affect the photo-realism of the final rendering.

On the other hand, generative adversarial networks (GANs) have recently shown great success in generating photo-realistic face images at high resolution [31]. Methods like StyleGAN [32] or StyleGAN2 [34] even provide high-level control over factors like pose or identity when trained on face images. However, these controlling factors are often entangled, and they are unknown prior to training. The control provided by these models does not allow the independent change of attributes like facial appearance, shape (e.g. length, width, roundness, etc.) or facial expression (e.g. raise eyebrows, open mouth, etc.). Although these methods have made significant progress in image quality, the provided control is not sufficient for graphics applications.

In short, we close the control-quality gap. However, we find that a naive conditional version of StyleGAN2 yields unsatisfactory results. We overcome this problem by rendering the geometric and photo-metric details of the FLAME mesh to an image and by using these as conditions instead. This design combination results in a generative 2D face model called GIF (Generative Interpretable Faces) that produces photo-realistic images with explicit control over face shape, head and jaw pose, expression, appearance, and illumination (Figure 1).

Finally we remark that the community lacks an auto-mated metric to effectively evaluate continuous conditional generative models. To this end we derive a quantitative score from a comparison-based perceptual study using an anonymous group of participants to facilitate future model comparisons.

In summary, our main contributions are 1) a generative 2D face model with FLAME [38] control, 2) use of FLAME renderings as conditioning for better association, 3) use of a texture consistency constraint to improve disentanglement, 4) providing a quantitative comparison mechanism.

2. Related Work

Generative 3D face models: Representing and manipulating human faces in 3D have a long standing history dating back almost five decades to the parametric 3D face model of Parke [45]. Blanz and Vetter [7] propose a 3D morphable model (3DMM), the first generative 3D face model that uses linear subspaces to model shape and appearance variations. This has given rise to a variety of 3D face models to model facial shape [9, 19, 47], shape and expression [1, 6, 8, 12, 49, 67], shape, expression and head pose [38], localized facial details [10, 43] and wrinkle details [27, 56]. However, renderings of these models do not reach photo-realism due to the lack of high-quality textures.

To overcome this, Saito et al. [53] introduce high-quality texture maps, and Slossberg et al. [58] and Gecer et al. [25] train GANs to synthesize textures with high-frequency details. While these works enhance the realism when being rendered by covering more texture details, they only model the face region (i.e. ignore hair, teeth, tongue, eyelids, eyes, etc.) required for photo-realistic rendering. While separate part-based generative models of hair [30, 52, 68], eyes [4], eyelids [5], ears [18], teeth [69], or tongue [29] exist, combining these into a complete realistic 3D face model remains an open problem.

Instead of explicitly modeling all face parts, Gecer et al. [24] use image-to-image translation to enhance the realism of images rendered from a 3D face mesh. Nagano et al. [42] generate dynamic textures that allow synthesizing expression dependent mouth interior and varying eye gaze. Despite significant progress of generative 3D face models [11, 21], they still lack photo-realism.

Our approach, in contrast, combines the semantic control of generative 3D models with the image synthesis ability of generative 2D models. This allows us to generate photo-realistic face images, including hair, eyes, teeth, etc. with explicit 3D face model controls.

Generative 2D face models: Early parametric 2D face models like Eigenfaces [57, 63], Fisherfaces [3], Active Shape Models [15], or Active Appearance Models [13] parametrize facial shape and appearance in images with linear spaces. Tensor faces [65] generalize these linear models to higher-order, generating face images with multiple

independent factors like identity, pose, or expression. Although these models provided some semantic control, they produced blurry and unrealistic images.

StyleGAN [32], a member of broad category of GAN [28] models, extends Progressive-GAN [31] by incorporating a style vector to gain partial control over the image generation, which is broadly missing in such models. However, the semantics of these controls are interpreted only post-training. Hence, it is possible that desired controls might not be present at all. InterFaceGAN [55] and StyleRig [59] aim to gain control over a pre-trained StyleGAN [32]. InterFaceGAN [55] identifies hyper-planes that separate positive and negative semantic attributes in a GAN's latent space. However, this requires categorical attributes for every kind of control, making it not suitable for a variety of aspects e.g. facial shape, expression etc. Further, many attributes might not be linearly separable. StyleRig [59] learns mappings between 3DMM parameters and the parameter vectors of each StyleGAN layer. StyleRig learns to edit StyleGAN parameters and thereby controls the generated image, with 3DMM parameters. The setup is mainly tailored towards face editing or face reenactment tasks. GIF, in contrast, provides full generative control over the image generation process (similar to regular GANs) but with semantically meaningful control over shape, pose, expression, appearance, lighting, and style.

CONFIG [37] leverages synthetic images to get ground truth control parameters and leverages real images to make the image generation look more realistic. However, generated images still lack photo-realism.

HoloGAN [44] randomly applies rigid transformations to learnt features during training, which provide explicit control over 3D rotations in the trained model. While this is feasible for global transformations, it remains unclear how to extend this to transformations of local parts or parameters like facial shape or expression. Similarly, IterGAN [23] also only models rigid rotation of generated objects using a GAN, but these rotations are restricted to 2D transformations.

Further works use variational autoencoders [51, 64] and flow-based methods [36] to generate images. These provide controllability, but do not reach the image quality of GANs.
Facial animation: A large body of work focuses on face editing or facial animation, which can be grouped into 3D model-based approaches (e.g. [26, 35, 39, 60, 61, 66]) or 3D model-free methods (e.g. [2, 48, 62, 70, 71])

Thies et al. [61] build a subject specific 3DMM from a video, reenact this model with expression parameters from another sequence, and blend the rendered mesh with the target video. Follow-up work use similar 3DMM-based retargeting techniques but replace the traditional rendering pipeline, or parts of it, with learnable components [35, 60]. Ververas and Zafeiriou [66] (Slider-GAN) and Geng et

al. [26] propose image-to-image translation models that, given an image of a particular subject, condition the face editing process on 3DMM parameters. Lombardi et al. [39] learn a subject-specific autoencoder of facial shape and appearance from high-quality multi-view images that allow animation and photo-realistic rendering. Like GIF, all these methods use explicit control of a pre-trained 3DMM or learn a 3D face model to manipulate or animate faces in images, but in contrast to GIF, they are unable to generate new identities.

Zakharov et al. [71] use image-to-image translation to animate a face image from 2D landmarks, ReenactGAN [70] and Recycle-GAN [2] transfer facial movement from a monocular video to a target person. Pumarola et al. [48] learn a GAN conditioned on facial action units for control over facial expressions. None of these methods provide explicit control over a 3D face representation.

All methods discussed above are task specific, i.e., they are dedicated towards manipulating or animating faces, while GIF, in contrast, is a generative 2D model that is able to generate new identities, and also provides control of a 3D face model. Further, most of the methods are trained on video data [2, 35, 39, 60, 61, 71], in contrast to GIF which is trained from static images only. Regardless, GIF can be used to generate facial animations.

3. Preliminaries

GANs and conditional GANs: GANs are a class of neural networks where a generator G and a discriminator D have opposing objectives. Namely, the discriminator estimates the probability of its input to be a generated sample, as opposed to a natural random sample from the training set, while the generator tries to make this task as hard as possible. This is extended in the case of conditional GANs [41]. The objective function in such a setting is given as

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p(x)}[\log D(x|c)] + \\ \mathbb{E}_{z \sim p(z)}[\log(1 - D(G(z|c)))], \quad (1)$$

where c is the conditioning variable. Although sound in an ideal setting, this formulation suffers a major drawback. Specifically under incomplete data regime, independent conditions tend to influence each other. In Section 4.2, we discuss this phenomenon in detail.

3.1. StyleGAN2

StyleGAN2 [33], a revised version of StyleGAN [31], produces photo-realistic face images at 1024×1024 resolution. Similar to StyleGAN, StyleGAN2, is controlled by a style vector $\mathbf{z}$. This vector is first transformed by a mapping network of 8 fully connected layers to $\mathbf{w}$, which then transforms the activations of the progressively growing resolution blocks using adaptive instance normalization (AdaIN)

layers. Although StyleGAN2 provides some high-level control, it still lacks explicit and semantically meaningful control. Our work addresses this shortcoming by distilling a conditional generative model out of StyleGAN2 and combining this with inputs from FLAME.

3.2. FLAME

FLAME is a publicly available 3D head model [38], $M(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}) : \mathbb{R}^{|\boldsymbol{\beta}| \times |\boldsymbol{\theta}| \times |\boldsymbol{\psi}|} \to \mathbb{R}^{N \times 3}$, which given parameters for facial shape $\boldsymbol{\beta} \in \mathbb{R}^{300}$, pose $\boldsymbol{\theta} \in \mathbb{R}^{15}$ (i.e. axis-angle rotations for global rotation and rotations around joints for neck, jaw, and eyeballs), and facial expression $\boldsymbol{\psi} \in \mathbb{R}^{100}$ outputs a mesh with $N = 5023$ vertices. We further transfer the appearance space of Basel Face Model [47] parametrized by $\boldsymbol{\alpha} \in \mathbb{R}^{|\boldsymbol{\alpha}|}$ into the FLAME's UV layout to augment it with a texture space. We use the same subset of FLAME parameters as RingNet [54], namely 6 pose coefficients for global rotation and jaw rotation, 100 shape, and 50 expression parameters, and we use 50 parameters for appearance. We use rendered FLAME meshes as the conditioning signal in GIF.

4. Method

Goal: Our goal is to learn a generative 2D face model controlled by a parametric 3D face model. Specifically, we seek a mapping $\text{GIF}(\Theta, \boldsymbol{\alpha}, \mathbf{l}, \mathbf{c}, \mathbf{s}) : \mathbb{R}^{156+50+27+3+512} \to \mathbb{R}^{P \times P \times 3}$, that given FLAME parameters $\Theta = \{\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}\} \in \mathbb{R}^{156}$, parameters for appearance $\boldsymbol{\alpha} \in \mathbb{R}^{50}$, spherical harmonics lighting $\mathbf{l} \in \mathbb{R}^{27}$, camera $\mathbf{c} \in \mathbb{R}^{3}$ (2D translation and isotropic scale of a weak-perspective camera), and style $\mathbf{s} \in \mathbb{R}^{512}$, generates an image of resolution $P \times P$.

Here, the FLAME parameters control all aspects related to the geometry, appearance and lighting parameters control the face color (i.e. skin tone, lighting, etc.), while the style vector $\mathbf{s}$ controls all factors that are not described by the FLAME geometry and appearance parameters, but are required to generate photo-realistic face images (e.g. hairstyle, background, etc.).

4.1. Training data

Our data set consists of Flickr images (FFHQ) introduced in StyleGAN [32] and their corresponding FLAME parameters, appearance parameters, lighting parameters, and camera parameters. We obtain these using DECA [22], a publicly available monocular 3D face regressor. In total, we use about 65,500 FFHQ images, paired with the corresponding parameters. The obtained FFHQ training parameters are available for research purposes.

4.2. Condition representation

Condition cross-talk: The vanilla conditional GAN formulation as described in Section 3 does not encode any semantic factorization of the conditional probability distributions

that might exist in the nature of the problem. Consider a situation where the true data depends upon two independent generating factors c_1 and c_2, i.e. the true generation process of our data is $x \sim P(x|c_1, c_2)$ where $P(x, c_1, c_2) = P(x|c_1, c_2) \cdot P(c_1) \cdot P(c_2)$. Ideally, given complete data (often infinite) and a perfect modeling paradigm, this factorization should emerge automatically. However, in practice, neither of these can be assumed. Hence, the representation of the condition highly influences the way it gets associated with the output. We refer to this phenomenon of independent conditions influencing each other as – *condition crosstalk*. Inductive bias introduced by condition representation in the context of conditional cross-talk is empirically evaluated in Section 5.

Pixel-aligned conditioning: Learning the rules of graphical projection (orthographic or perspective) and the notion of occlusion as part of the generator is wasteful if explicit 3D geometry information is present, as it approximates classical rendering operations, which can be done learning-free and are already part of several software packages (e.g. [40, 50]). Hence, we provide the generator with explicit knowledge of the 3D geometry by conditioning it with renderings from a classical renderer. This makes pixel-localized association between the FLAME conditioning signal and the generated image possible. We find that, although a vanilla conditional GAN achieves comparable image quality, GIF learns to better obey the given condition.

We condition GIF on two renderings, one provides pixel-wise color information (referred to as texture rendering), and the other provides information about geometry (referred to as normal rendering). Normal renderings are obtained by rendering the mesh with a color-coded map of the surface normals $\mathbf{n} = \mathbf{N}(M(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}))$.

Both renderings use a scaled orthographic projection with the camera parameters provided with each training image. For the color rendering, we use the provided inferred lighting and appearance parameter. The texture and the normal renderings are concatenated along the color channel and used to condition the generator as shown in Figure 2. As demonstrated in Section 5, this conditioning mechanism helps reduce condition cross-talk.

4.3. GIF architecture

The model architecture of GIF is based on StyleGAN2 [33], with several key modifications, discussed as follows. An overview is shown in Figure 2.

Style embedding: Rendering the textured FLAME mesh does not consider hair, mouth cavity (i.e., teeth or tongue), and fine-scale details like wrinkles and pores. Generating a realistic face image, however, requires these factors to be considered. Hence we introduce a style vector $\mathbf{s} \in \mathbb{R}^{512}$ to model these factors. Note that original StyleGAN2 has a similar vector $\mathbf{z}$ with the same dimensionality, however,
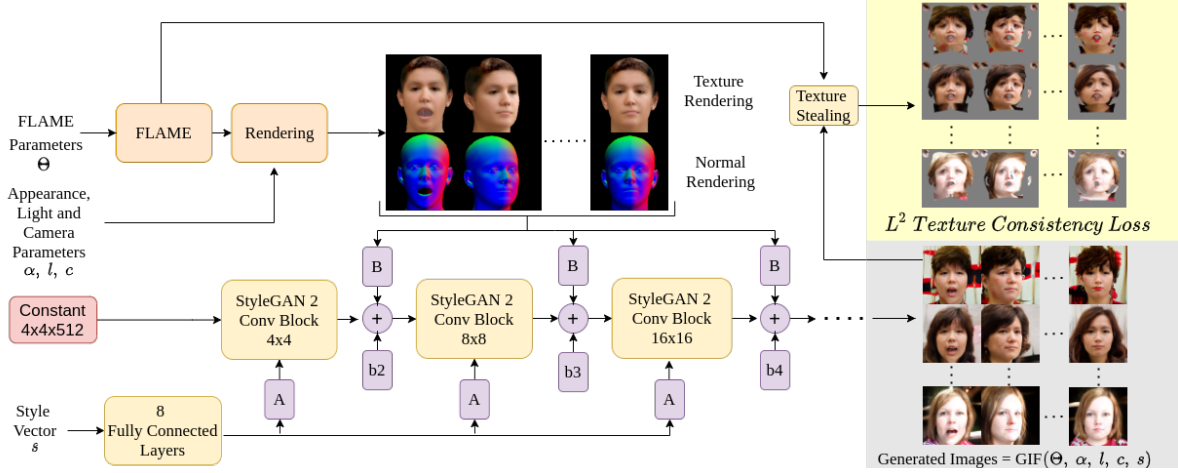
Figure 2: Our generator architecture is based on StyleGAN2 [33], A is a learned affine transform, and B stands for per-channel scaling. We make several key changes, such as introducing 3D model generated condition through the noise injection channels and introduce texture consistency loss. We refer to the process of projecting the generated image onto the FLAME mesh to obtain an incomplete texture map as *texture stealing*.

instead of drawing random samples from a standard normal $\mathcal{N}(0, I)$ distribution (as common in GANs), we assign a random but unique vector for every image. This is motivated by the key observation that in the FFHQ dataset [32], each identity mostly occurs only once and mostly has a unique background. Hence, if we use a dedicated vector for each image using, e.g., an embedding layer, we will encode an inductive bias for this vector to capture background and appearance specific information.

Noise channel conditioning: StyleGAN and StyleGAN2 insert random noise images at each resolution level into the generator, which mainly contributes to local texture changes. We replace this random noise by the concatenated textured and normal renderings from the FLAME model, and insert scaled versions of these renderings at different resolutions into the generator. This is motivated by the observation that varying FLAME parameters, and therefore varying FLAME renderings, should have direct, pixel-aligned influence on the generated images.

4.4. Texture consistency

Since GIF's generation is based on an underlying 3D model, we can further constrain the generator by introducing a texture consistency loss optimized during training. We first generate a set of new FLAME parameters by randomly interpolating between the parameters within a mini batch. Next, we generate the corresponding images with the same style embedding **s**, appearance α and lighting parameters **l** by an additional forward pass through the model. Finally, the corresponding FLAME meshes are projected onto the generated images to get a partial texture map (also referred to as 'texture stealing'). To enforce pixel-wise consistency,

we apply an L^2 loss on the difference between pairs of texture maps, considering only pixels for which the corresponding 3D point on the mesh surface is visible in both the generated images. We find that this texture consistency loss improves the parameter association (see Section 5).

5. Experiments

5.1. Qualitative evaluation

Condition influence: As described in Section 4, GIF is parametrized by FLAME parameters $\Theta = \{\beta, \theta, \psi\}$, appearance parameters α, lighting parameters **l**, camera parameters **c** and a style vector **s**. Figure 3 shows the influence of each individual set of parameters by progressively exchanging one type of parameter in each row. The top and bottom rows show GIF generated images for two sets of parameters, randomly chosen from the training data.

Exchanging style (row 2) most noticeably changes hair style, clothing color, and the background. Shape (row 3) is strongly associated to the person's identity among other factors. The expression parameters (row 4) control the facial expression, best visible around the mouth and cheeks. The change in pose parameters (row 5) affects the orientation of the head (i.e. head pose) and the extent of the mouth opening (jaw pose). Finally, appearance (row 6) and lighting (row 7) change the skin color and the lighting specularity.

Random sampling: To further evaluate GIF qualitatively, we sample FLAME parameters, appearance parameters, lighting parameters, and style embeddings and generate random images, shown in Figure 4. For shape, expression, and appearance parameters, we sample parameters of the first three principal components from a standard normal distri-

GIF($\beta_1, \theta_1, \psi_1, \alpha_1, l_1, c_1, s_1$)

GIF($\beta_1, \theta_1, \psi_1, \alpha_1, l_1, c_1, \mathbf{s_2}$)

$GIF(\beta_2, \theta_1, \psi_1, \alpha_1, l_1, c_1, \mathbf{s_2}$)

GIF($\beta_2, \theta_1, \psi_2, \alpha_1, l_1, c_1, \mathbf{s_2}$)

GIF($\beta_2, \theta_2, \psi_2, \alpha_1, l_1, c_2, \mathbf{s_2}$)

GIF($\beta_2, \theta_2, \psi_2, \alpha_2, l_1, c_2, \mathbf{s_2}$)

GIF($\beta_2, \theta_2, \psi_2, \alpha_2, l_2, c_2, \mathbf{s_2}$)

Figure 3: Impact of individual parameters, when being exchanged between two different generated images one at a time. From top to bottom we progressively exchange style, shape, expression, head and jaw pose, appearance, and lighting of the two parameter sets. We progressively change the color of the parameter symbol that is effected in every row to red.

bution and keep all other parameters at zero. For pose, we sample from a uniform distribution in $[-\pi/8, +\pi/8]$ (head pose) for rotation around the y-axis, and $[0, +\pi/12]$ (jaw pose) around the x-axis. For lighting parameters and style embeddings, we choose random samples from the set of training parameters. Figure 4 shows that GIF produces photo-realistic images of different identities with a large variation in shape, pose, expression, skin color, and age. Figure 1 further shows rendered FLAME meshes for generated images, demonstrating that GIF generated images are well associated with the FLAME parameters.

Speech driven animation: As GIF uses FLAME's parametric control, it can directly be combined with existing FLAME-based application methods such as VOCA [17], which animates a face template in FLAME mesh topology from speech. For this, we run VOCA for a speech sequence, fit FLAME to the resulting meshes, and use these parame-

	Vector cond.	No Texture interpolation	Normal rend. conditioning	Texture rend. conditioning
GIF	89.4%	51.1%	55.8%	51.7%

Table 1: AMT ablation experiment. Preference percentage of GIF generated images over ablated models and vector conditioning model. Participants were instructed to pay particular attention to shape, pose, and expression and ignore image quality.

ters to drive GIF for different appearance embeddings (see Figure 5). For more qualitative results and the full animation sequence, see the supplementary video.

5.2. Quantitative evaluation

We conduct two Amazon Mechanical Turk (AMT) studies to quantitatively evaluate i) the effects of ablating individual model parts, and ii) the disentanglement of geometry and style. We compare GIF in total with 4 different ablated versions, namely *vector conditioning*, *no texture interpolation*, *normal rendering*, and *texture rendering conditioning*. For the *vector conditioning* model, we directly provide the FLAME parameters as a 236 dimensional real valued vector. In the *no texture interpolation* model, we drop the texture consistency during training. The *normal rendering* and *texture rendering conditioning* models condition the generator and discriminator only on the normal rendering or texture rendering, respectively, while removing the other rendering. The supplementary video shows examples for both user studies.

Ablation experiment: Participants see three images, a reference image in the center, which shows a rendered FLAME mesh, and two generated images to the left and right in random order. Both images are generated from the same set of parameters, one using GIF, another an ablated GIF model. Participants then select the generated image that corresponds best with the reference image. Table 1 shows that with the texture consistency loss, normal rendering, and texture rendering conditioning, GIF performs slightly better than without each of them. Participants tend to select GIF generated results over a vanilla conditional StyleGAN2 (please refer to our supplementary material for details on the architecture). Furthermore, Figure 2 quantitatively evaluates the image quality with FID scores, indicating that all models produce similar high-quality results.

Geometry-style disentanglement: In this experiment, we study the entanglement between the style vector and FLAME parameters. We find qualitatively that the style vector mostly controls aspects of the image that are not influenced by FLAME parameters like background, hairstyle, etc. (see Figure 3). To evaluate this quantitatively, we randomly pair style vectors and FLAME parameters and

Figure 4: Images obtained by randomly sampling FLAME, appearance parameters, style parameters, and lighting parameters. Specifically, for shape, expression, and appearance parameters, we sample parameters of the first three principal components from a standard normal distribution and keep all other parameters at zero. We sample pose from a uniform distribution in $[-\pi/8, +\pi/8]$ (head pose) for rotation around the y-axis, and $[0, +\pi/12]$ (jaw pose) around the x-axis
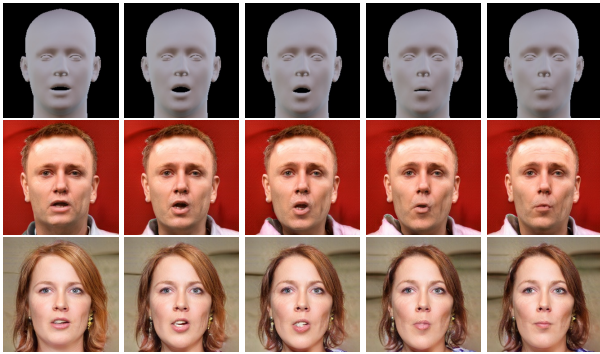

Figure 5: Combination of GIF and an existing speech-driven facial animation method by generating face images for FLAME parameters obtained from VOCA [17]. Sample frames to highlight jaw pose variation. Please see the supplementary video for the full animation.

GIF	Vector cond.	No Texture interpolation	Normal rend. conditioning	Texture rend. conditioning
8.94	10.34	11.71	9.89	11.28

Table 2: FID scores of images generated by GIF and ablated models (lower is better). Note that this score only evaluates image quality and does not judge how well the models obey the underlying FLAME conditions.

conduct a perceptual study with AMT. Participants see a rendered FLAME image and GIF generated images with the same FLAME parameters but with a variety of different style vectors. Participants then rate the similarity of the generated image to shape, pose and expression of the FLAME rendering on a standard 5-Point Likert scale (i.e.

1: Strongly Disagree, 2. Disagree, 3. Neither agree nor disagree, 4. Agree, 5. Strongly Agree). We use 10 randomized styles $\mathbf{s}$ and 500 random FLAME parameters totaling to 5000 images. We find that the majority of the participants agree that generated images and FLAME rendering are similar, irrespective of the style (see Figure 6).

Re-inference error: In the spirit of DiscoFaceGAN [20], we run DECA [22] on the generated images and compute the Root Mean Square Error (RMSE) between the FLAME face vertices of the input parameters and the inferred parameters as reported in Table 3. We generate a population of 1000 images by randomly sampling one of the shape, pose and expression latent spaces while holding the rest of the generating factors to their neutral. Thus we find an association error for individual factors.

Model	Shape	Expression	Pose
Vector cond.	3.43 mm	23.05 mm	29.69 mm
GIF	**3.02** mm	**5.00** mm	**5.61** mm

Table 3: To evaluate shape error, we generate 1024 random faces from GIF and re-infer their shape using DECA [22]. Using FLAME, we compute face region vertices twice: once with GIF's input condition and once with re-inferred parameters. Finally, the RMSE between these face region vertices are computed. The process is repeated for expression and pose.

6. Discussion

GIF is trained on the FFHQ data-set and hence inherits some of its limitations, e.g. the images are roughly
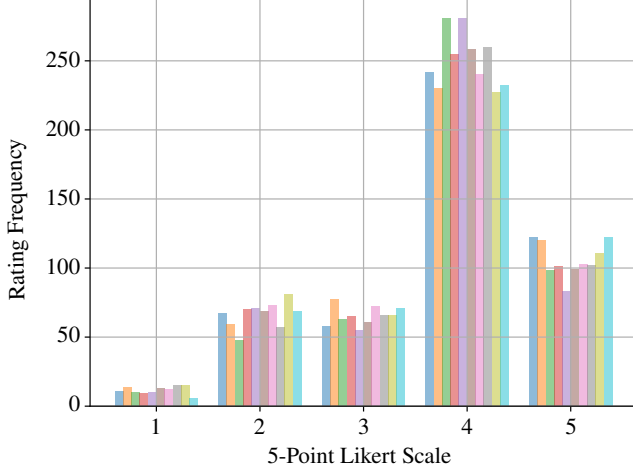
Figure 6: Preference frequency of styles on a 5-Point Likert Scale. Note that almost all style vectors, represented with different colors here get a similar distribution of likeness ratings indicating that they do not influence the perceived FLAME conditioning.

eye-centered. Although StyleGAN2 [34] has to some extent addressed this issue (among others), but failed to do so completely. Hence, rotations of the head look like an eye-centered rotation, which involves a combination of 3D rotation and translation as opposed to a pure neck centered rotation. Adapting GIF to use an architecture other than StyleGAN2 or training it on a different data set to improve image quality is subject to future work.

As FLAME renderings for conditioning GIF must be similarly eye-centered as the FFHQ training data. We compute a suitable camera parameter from given FLAME parameters so that the eyes are located roughly at the centre of the image plane with a fixed distance between the eyes. However, for profile view poses, this can not be met without an extreme zoomed in view. This often causes severe artifacts. Please see the supplementary material for examples.

GIF requires a statistical 3D model, and a way to associate its parameters to a large data set of high-quality images. While 'objects' like human bodies [46] or animals [72] potentially fulfill these requirements, it remains unclear how to apply GIF to general object categories.

Faulty image to 3D model associations stemming from the parameter inference method potentially degrade GIF's generation quality. One example is the ambiguity of lighting and appearance, which causes most color variations in the training data to be described by lighting variation rather than by appearance variation. GIF inherits these errors.

Finally, as GIF is solely trained from static images without multiple images per subject, generating images with varying FLAME parameters is not temporally consistent. As several unconditioned parts are only loosely correlated

or uncorrelated to the condition (e.g. hair, mouth cavity, background, etc.), this results in jittery video sequences. Training or refining GIF on temporal data with additional temporal constraints during training remains subject to future work.

7. Conclusion

We present GIF, a generative 2D face model, with high realism and with explicit control from FLAME, a statistical 3D face model. Given a data set of approximately 65,500 high-quality face images with associated FLAME model parameters for shape, global pose, jaw pose, expression, and appearance parameters, GIF learns to generate realistic face images that associate with them. Our key insight is that conditioning a generator network on explicit information rendered from a 3D face model allows us to decouple shape, pose, and expression variations within the trained model. Given a set of FLAME parameters associated with an image, we render the corresponding FLAME mesh twice, once with color-coded normal details, once with an inferred texture, and insert these as condition to the generator. We further add a loss that enforces consistency in texture for the reconstruction of different FLAME parameters for the same appearance embedding. This encourages the network during training to disentangle appearance and FLAME parameters, and provides us with better temporal consistency when generating frames of FLAME sequences. Finally we devise a comparison-based perceptual study to evaluate continuous conditional generative models quantitatively.

8. Acknowledgements

9. Disclosure

MJB has received research gift funds from Intel, Nvidia, Adobe, Facebook, and Amazon. While MJB is a part-time employee of Amazon, his research was performed solely at, and funded solely by, MPI. MJB has financial interests in Amazon and Meshcapade GmbH. PG has received funding from Amazon Web Services for using their Machine Learning Services and GPU Instances. AR's research was performed solely at, and funded solely by, MPI.

References

[1] B. Amberg, R. Knothe, and T. Vetter. Expression invariant 3D face recognition with a morphable model. In *Interna-*

tional Conference on Automatic Face & Gesture Recognition (FG), pages 1–6, 2008. 2

[2] A. Bansal, S. Ma, D. Ramanan, and Y. Sheikh. Recycle-GAN: Unsupervised video retargeting. In *European Conference on Computer Vision (ECCV)*, pages 119–135, 2018. 3

[3] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 19(7):711–720, 1997. 2

[4] P. Bérard, D. Bradley, M. Nitti, T. Beeler, and M. Gross. High-quality capture of eyes. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 33(6):1–12, 2014. 2

[5] A. Bermano, T. Beeler, Y. Kozlov, D. Bradley, B. Bickel, and M. Gross. Detailed spatio-temporal reconstruction of eyelids. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 34(4):1–11, 2015. 2

[6] V. Blanz, C. Basso, T. Poggio, and T. Vetter. Reanimating faces in images and video. In *Computer graphics forum*, volume 22, pages 641–650, 2003. 2

[7] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, pages 187–194, 1999. 2

[8] T. Bolkart and S. Wuhrer. A groupwise multilinear correspondence optimization for 3D faces. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3604–3612, 2015. 2

[9] J. Booth, A. Roussos, A. Ponniah, D. Dunaway, and S. Zafeiriou. Large scale 3D morphable models. *International Journal of Computer Vision (IJCV)*, 126(2-4):233–254, 2018. 2

[10] A. Brunton, T. Bolkart, and S. Wuhrer. Multilinear wavelets: A statistical shape space for human faces. In *European Conference on Computer Vision (ECCV)*, pages 297–312, 2014. 2

[11] A. Brunton, A. Salazar, T. Bolkart, and S. Wuhrer. Review of statistical shape spaces for 3D data with comparative analysis for human faces. *Computer Vision and Image Understanding (CVIU)*, 128:1–17, 2014. 2

[12] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. FaceWarehouse: A 3D facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2013. 2

[13] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *European Conference on Computer Vision (ECCV)*, pages 484–498, 1998. 2

[14] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 23(6):681–685, 2001. 2

[15] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models-their training and application. *Computer Vision and Image Understanding (CVIU)*, 61(1):38–59, 1995. 2

[16] I. Craw and P. Cameron. Parameterising images for recognition and reconstruction. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 367–370. 1991. 2

[17] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. Black. Capture, learning, and synthesis of 3D speaking styles. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10101–10111, 2019. 6, 7

[18] H. Dai, N. Pears, and W. Smith. A data-augmented 3D morphable model of the ear. In *International Conference on Automatic Face & Gesture Recognition (FG)*, pages 404–408, 2018. 2

[19] H. Dai, N. Pears, W. A. Smith, and C. Duncan. A 3D morphable model of craniofacial shape and texture variation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3085–3093, 2017. 2

[20] Y. Deng et al. Disentangled and controllable face image generation via 3D imitative-contrastive learning. In *CVPR*, pages 5154–5163, 2020. 7

[21] B. Egger, W. A. Smith, A. Tewari, S. Wuhrer, M. Zollhoefer, T. Beeler, F. Bernard, T. Bolkart, A. Kortylewski, S. Romdhani, et al. 3D morphable face models–past, present and future. *ACM Transactions on Graphics (TOG)*, 39(5), 2020. 2

[22] Y. Feng, H. Feng, M. J. Black, and T. Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. *CoRR*, 2020. 4, 7

[23] Y. Galama and T. Mensink. IterGANs: Iterative GANs to learn and control 3D object transformation. *Computer Vision and Image Understanding (CVIU)*, 189:102803, 2019. 3

[24] B. Gecer, B. Bhattarai, J. Kittler, and T.-K. Kim. Semi-supervised adversarial learning to generate photorealistic face images of new identities from 3D morphable model. In *European Conference on Computer Vision (ECCV)*, pages 217–234, 2018. 2

[25] B. Gecer, S. Ploumpis, I. Kotsia, and S. Zafeiriou. GANFIT: Generative adversarial network fitting for high fidelity 3D face reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1155–1164, 2019. 2

[26] Z. Geng, C. Cao, and S. Tulyakov. 3D guided fine-grained face manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9821–9830, 2019. 3

[27] A. Golovinskiy, W. Matusik, H. Pfister, S. Rusinkiewicz, and T. Funkhouser. A statistical model for synthesis of detailed facial geometry. *ACM Transactions on Graphics (TOG)*, 25(3):1025–1034, 2006. 2

[28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014. 3

[29] A. Hewer, S. Wuhrer, I. Steiner, and K. Richmond. A multilinear tongue model derived from speech related MRI data of the human vocal tract. *Computer Speech & Language*, 51:68–92, 2018. 2

[30] L. Hu, D. Bradley, H. Li, and T. Beeler. Simulation-ready hair capture. In *Computer Graphics Forum*, volume 36, pages 281–294, 2017. 2

[31] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. 2018. 2, 3

[32] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019. 2, 3, 4, 5

[33] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of Style-GAN. *CoRR*, abs/1912.04958, 2019. 3, 4, 5

[34] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of Style-GAN. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8110–8119, 2020. 2, 8

[35] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Nießner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt. Deep video portraits. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018. 3

[36] D. P. Kingma and P. Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pages 10215–10224, 2018. 3

[37] M. Kowalski, S. J. Garbin, V. Estellers, T. Baltrušaitis, M. Johnson, and J. Shotton. CONFIG: Controllable neural face image generation. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[38] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 1, 2, 4

[39] S. Lombardi, J. Saragih, T. Simon, and Y. Sheikh. Deep appearance models for face rendering. *ACM Transactions on Graphics (TOG)*, 37(4):1–13, 2018. 3

[40] M. M. Loper and M. J. Black. OpenDR: An approximate differentiable renderer. In *European Conference on Computer Vision (ECCV)*, volume 8695, pages 154–169, 2014. 4

[41] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 3

[42] K. Nagano, J. Seo, J. Xing, L. Wei, Z. Li, S. Saito, A. Agarwal, J. Fursund, H. Li, R. Roberts, et al. paGAN: real-time avatars using dynamic textures. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 37(6):258–1, 2018. 2

[43] D. R. Neog, J. L. Cardoso, A. Ranjan, and D. K. Pai. Interactive gaze driven animation of the eye region. In *Proceedings of the 21st International Conference on Web3D Technology*, pages 51–59, 2016. 2

[44] T. Nguyen-Phuoc, C. Li, L. Theis, C. Richardt, and Y.-L. Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 7588–7597, 2019. 3

[45] F. I. Parke. A parametric model for human faces. Technical report, UTAH UNIV SALT LAKE CITY DEPT OF COMPUTER SCIENCE, 1974. 2

[46] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 8

[47] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3D face model for pose and illumination invariant face recognition. In *International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301, 2009. 2, 4

[48] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *European Conference on Computer Vision (ECCV)*, pages 818–833, 2018. 3

[49] A. Ranjan, T. Bolkart, S. Sanyal, and M. J. Black. Generating 3D faces using convolutional mesh autoencoders. In *European Conference on Computer Vision (ECCV)*, pages 725–741, 2018. 2

[50] N. Ravi, J. Reizenstein, D. Novotny, T. Gordon, W.-Y. Lo, J. Johnson, and G. Gkioxari. Pytorch3d. https://github.com/facebookresearch/pytorch3d, 2020. 4

[51] A. Razavi, A. van den Oord, and O. Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *Advances in Neural Information Processing Systems*, pages 14837–14847, 2019. 3

[52] S. Saito, L. Hu, C. Ma, H. Ibayashi, L. Luo, and H. Li. 3D hair synthesis using volumetric variational autoencoders. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 37(6):1–12, 2018. 2

[53] S. Saito, L. Wei, L. Hu, K. Nagano, and H. Li. Photorealistic facial texture inference using deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5144–5153, 2017. 2

[54] S. Sanyal, T. Bolkart, H. Feng, and M. J. Black. Learning to regress 3D face shape and expression from an image without 3d supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7763–7772, 2019. 4

[55] Y. Shen, J. Gu, X. Tang, and B. Zhou. Interpreting the latent space of GANs for semantic face editing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9243–9252, 2020. 3

[56] I.-K. Shin, A. C. Öztireli, H.-J. Kim, T. Beeler, M. Gross, and S.-M. Choi. Extraction and transfer of facial expression wrinkles for facial performance enhancement. In *Pacific Conference on Computer Graphics and Applications*, pages 113–118, 2014. 2

[57] L. Sirovich and M. Kirby. Low-dimensional procedure for the characterization of human faces. *Optical Society of America A*, 4(3):519–524, 1987. 2

[58] R. Slossberg, G. Shamai, and R. Kimmel. High quality facial surface and texture synthesis via generative adversarial networks. In *European Conference on Computer Vision Workshops (ECCV-W)*, 2018. 2

[59] A. Tewari, M. Elgharib, G. Bharaj, F. Bernard, H.-P. Seidel, P. Pérez, M. Zollhofer, and C. Theobalt. StyleRig: Rigging StyleGAN for 3D control over portrait images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6142–6151, 2020. 3

[60] J. Thies, M. Zollhöfer, and M. Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 3

[61] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Niessner. Face2Face: Real-time face capture and reenactment of RGB videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2387–2395, 2016. 3

[62] S. Tripathy, J. Kannala, and E. Rahtu. ICface: Interpretable and controllable face reenactment using GANs. In *Winter Conference on Applications of Computer Vision (WACV)*, pages 3385–3394, 2020. 3

[63] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991. 2

[64] A. van den Oord, O. Vinyals, et al. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, pages 6306–6315, 2017. 3

[65] M. A. O. Vasilescu and D. Terzopoulos. Multilinear analysis of image ensembles: Tensorfaces. In *European Conference on Computer Vision (ECCV)*, pages 447–460, 2002. 2

[66] E. Ververas and S. Zafeiriou. Slidergan: Synthesizing expressive face images by sliding 3D blendshape parameters. *International Journal of Computer Vision (IJCV)*, pages 1–22, 2020. 3

[67] D. Vlasic, M. Brand, H. Pfister, and J. Popović. Face transfer with multilinear models. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 24(3):426–433, 2005. 2

[68] L. Wei, L. Hu, V. Kim, E. Yumer, and H. Li. Real-time hair rendering using sequential adversarial networks. In *European Conference on Computer Vision (ECCV)*, pages 99–116, 2018. 2

[69] C. Wu, D. Bradley, P. Garrido, M. Zollhöfer, C. Theobalt, M. H. Gross, and T. Beeler. Model-based teeth reconstruction. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 35(6):220–1, 2016. 2

[70] W. Wu, Y. Zhang, C. Li, C. Qian, and C. Change Loy. Reenactgan: Learning to reenact faces via boundary transfer. In *European Conference on Computer Vision (ECCV)*, pages 603–619, 2018. 3

[71] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9459–9468, 2019. 3

[72] S. Zuffi, A. Kanazawa, T. Berger-Wolf, and M. J. Black. Three-D safari: Learning to estimate zebra pose, shape, and texture from images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5359–5368, 2019. 8
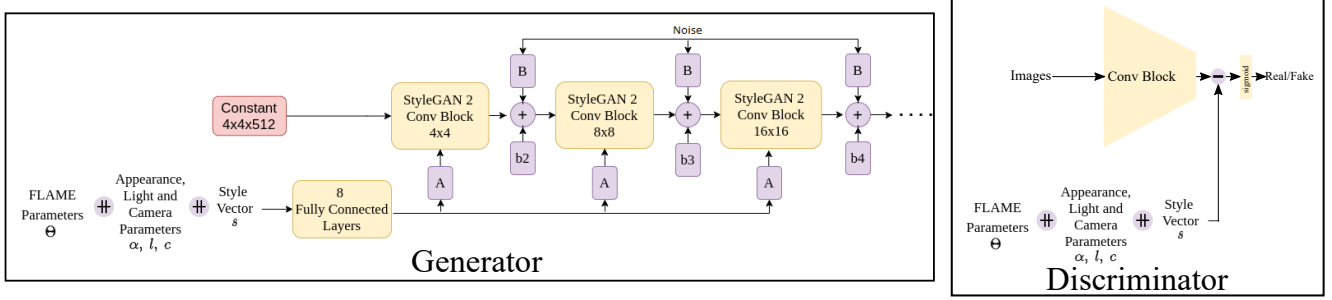
Figure 7: Architecture of Vector conditioning model. Here ⊞ represents a concatenation.

A. Vector conditioning architecture

Here we describe the model architecture of the vector condition model, used as one of the baseline models in Section 5.2 of the main paper. For this model, we pass the vector values conditioning parameters FLAME $(\beta, \theta, \psi,)$, appearance (α) and lighting (l) as a 236 dimensional vector through the dimensions of style vector of the original StyleGAN2 architecture as shown in Figure 7. We further input the same conditioning vector to the discriminator at the last fully connected layer of the discriminator by subtracting it from the last layer activation.

profile views however it is unclear how the centering within the training data was achieved. The centering strategy used in GIF causes a zoom in for profile views, effectively cropping parts of the face, and hence the generator struggles to generate realistic images as shown in Figure 8.
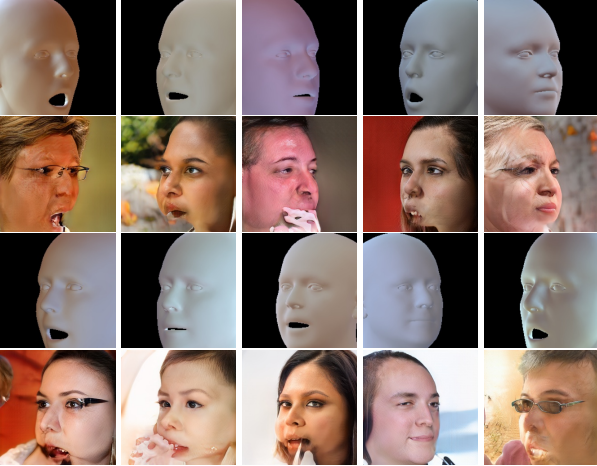

Figure 8: In order for the eyes to be places at a given pixel location a profile view causes the face image to be highly zoomed in. This causes the generated images to become unrealistic.

B. Image centering for extreme rotations

As discussed in Section 6 of the main paper, GIF produces artifacts for extreme head poses close to profile view. This is due to the pixel alignment of the FLAME renderings and the generated images, which requires the images to be similarly eye-centered as the FFHQ training data. For