GIF: Generative Interpretable Faces

Partha Ghosh¹ Pravir Singh Gupta*² Roy Uziel*³
Anurag Ranjan¹ Michael Black¹ Timo Bolkart¹

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,anurag.ranjan,black,timo.bolkart}@tue.mpg.de pravir@tamu.edu uzielr@post.bgu.ac.il

Abstract

Photo-realistic visualization and animation of expressive human faces have been a long standing challenge. On one end of the spectrum, 3D face modeling methods provide parametric control but tend to generate unrealistic images, while on the other end, 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. Trained GANs without pre-defined control, however, may entangle factors that are hard to undo later. To guarantee some disentanglement that provides us with desired kinds of control, we train our generative model conditioned on pre-defined control parameters. Specifically, we condition StyleGAN2 on FLAME, a generative 3D face model. However, we found out that a naive conditioning on FLAME parameters yields rather unsatisfactory results. Instead we render out geometry and photo-metric details of the FLAME mesh and use these for conditioning instead. This gives us a generative 2D face model named GIF (Generative Interpretable Faces) that shares FLAME’s parametric control. Given FLAME parameters for shape, pose, and expressions, parameters for appearance and lighting, and an additional style vector, GIF outputs photo-realistic face images. To evaluate how well GIF follows its conditioning and the impact of different design choices, we perform a perceptual study. The code and trained model are publicly available for research purposes at https://github.com/ParthaEth/GIF.

*Equal contribution.
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 [17, 16, 14, 15, 62] or higher-order tensor generalizations [64]. Although they provide some semantic control, these methods use linear transformations in 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 [8] introduced a statistical 3D morphable model, of facial shape and appearance. Such, a statistical face model (e.g. [13, 46, 38]) is used to manipulate shape, expression, or pose of the facial 3D geometry. This is then combined with a texture map (e.g. [52]) 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 to independently change attributes like facial appearance, shape (e.g. length, width, roundness, etc.) or facial expression (e.g. raise eyebrows, open mouth, etc.) without changing other factors too. Although these methods have made significant progress in image quality, the provided control is not sufficient for graphics applications.

In short, the discipline of generating face images reveals a gap between low quality face images with explicit control from a 3D face model on one side of the spectrum, and high-quality images from 2D generative models without explicit control on the other side. We close this gap by incorporating 3D geometry from FLAME [38] (a publicly available statistical 3D face model) as a supervised condition to a high quality generative 2D model StyleGAN2 [33]. However, we found out that a naive conditional version of StyleGAN2 yields rather unsatisfactory results. We overcome this problem by rendering out geometric and photo-geometric details of the FLAME mesh and by using them 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).

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 texture consistency constraint to improve disentanglement.

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 [44]. Blanz and Vetter [8] 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 [10, 46, 20], shape and expression [1, 9, 7, 13, 48, 66], shape, expression and head pose [38], localized facial details [11] and wrinkle details [27, 55]. However, renderings of these models do not reach photo-realism due to the lack of high-quality textures.

To overcome this, Saito et al. [52] introduce high-quality texture maps, and Slossberg et al. [57] 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. ignores hair, teeth, tongue, eyelids, eyes, etc.) required for photo-realistic rendering. While separate part-based generative models of hair [30, 51, 67], eyes [5], eyelids [6], ears [19], teeth [68], 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 [12, 22], 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 [56, 62], Fisherfaces [4], Active Shape Models [16], or Active Appearance Models [14] parametrize facial shape and appearance in images with linear spaces. Tensor faces [64] generalize these linear models to higher-order, generating face images with multiple independent factors like identity, pose, or expression. Al-
though these models provided some semantic control, they produced blurry and unrealistic images.

GANs [28] and its variants, e.g. Progressive-GAN [31] on the other side, generate photo-realistic face images but lack semantic control. StyleGAN [32] extends Progressive-GAN [31] by incorporating a style vector to recover control over the image generation. As the semantics of the effects of the style vector are interpreted post-training, it is possible that desired controls might not be present at all. This drastically limits the potential for applications. InterFaceGAN [54] and StyleRig [58] aim to gain control over a pre-trained StyleGAN [32]. InterFaceGAN [54] identifies hyper-planes that separate positive and negative semantic attributes in the latent space of a GAN for semantic controls. However, this requires categorical attributes for every kind of control. These are not particularly suitable representation for many model factors e.g. facial shape. Further, many attributes might not be linearly separable in the latent space without any supervision. StyleRig [58] learns mappings between 3DMM parameters and the parameter vectors of each StyleGAN layer. Given a set of StyleGAN parameter vectors and pose, expression, and lighting parameters, StyleRig outputs a set of parameters, which is fed to StyleGAN to generate a photo-realistic face image. StyleRig learns to edit StyleGAN parameters, and hence a particular StyleGAN generated image, with 3DMM control, which is tailored to face editing or face reenactment tasks. GIF in contrast provides full generative control over the image generation process (similar to regular GANs) but with semantic meaningful control over shape, pose, expression, appearance, lighting, and style. Further, StyleRig inherits inductive biases of the underlying unconditional GAN, while GIF enforces disentanglement by conditioning on the pre-defined control parameters during GAN training. In summary while StyleRig tries to gain control over StyleGAN’s generation process post hoc, GIF directly modifies the training process such that the controls emerge naturally. CONFIG [37] combines synthetic images of faces along with unlabeled real ones to gain control over the image generation process. During training, CONFIG tries to lower the representation gap in the shared factorized embedding space of real and synthetic encoders. For inference, swapping sets of latent parameters then only influences desired aspects of the generated image. However, supposedly due to the large domain gap between synthetic and real images, generated images lack photo-realism. HoloGAN [43] randomly applies rigid transformations to learnt features during training, which provides 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 [50, 63] and flow-based methods [36] to generate images. These provides controllability, but does 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, 59, 60, 65]) or 3D model-free methods (e.g. [2, 47, 61, 69, 70]).

Thies et al. [60] 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 re-targeting techniques but replace the traditional rendering pipeline or parts of it with learnable components [35, 59]. Ververas and Zafeiriou [65] (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 allows 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. [70] use image-to-image translation to animate a face image from 2D landmarks, ReenactGAN [69] and Recycle-GAN [2] transfer facial movement from a monocular video to a target person. Pumarola et al. [47] 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 it provides control of a 3D face model. Further, most of the methods are trained on video data [2, 60, 35, 59, 39, 70], 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: Generative Adversarial Networks (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]. Here the discriminator estimates the probabilities of the above-mentioned quantities given the conditions. The objective function in such a
setting is given as
\[
\min_G \max_D V(D, G) = \mathbb{E}_{z \sim p(z)}[\log D(x|c)] + \mathbb{E}_{z \sim p(z)}[\log(1 - D(G(z|c)))],
\]
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 \(\times\) 1024 resolution. Similar to StyleGAN, StyleGAN2, is controlled by a style vector \(z\). This vector is first transformed by a mapping network of 8 fully connected layers to \(w\), which then transforms the activations of the progressively growing resolution blocks using adaptive instance normalization (AdaIN) layers. StyleGAN2 achieves significantly better visual image quality as compared to its predecessors. 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. This not only helps to reduce the gap of photo-realism and controlled generation of face images but also sheds light on inductive bias of the condition-representation on the disentanglement of conditional GANs.

3.2. FLAME

FLAME is a publicly available 3D head model [38], \(M(\beta, \theta, \psi) : \mathbb{R}^{[\beta] \times [\theta] \times [\psi]} \rightarrow \mathbb{R}^{N \times 3}\), which given parameters for facial shape \(\beta \in \mathbb{R}^{300}\), pose \(\theta \in \mathbb{R}^{15}\) (i.e. axis-angle rotations for global rotation and rotations around joints for neck, jaw, and eyeballs), and facial expression \(\psi \in \mathbb{R}^{100}\) outputs a mesh with \(N = 5023\) vertices. We further transfer the appearance space of Basel Face Model [46] parametrized by \(\alpha \in \mathbb{R}^{[\alpha]}\) into the FLAME's UV layout to augment it with a texture space. We use the same subset of FLAME parameters as done by RingNet [53], 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. This way we achieve both semantic control and photo-realism.

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 \(\text{GIF}(\Theta, \alpha, l, c, s) : \mathbb{R}^{156 + 50 + 27 + 3 + 512} \rightarrow \mathbb{R}^{P \times P \times 3}\), that given FLAME parameters \(\Theta = \{\beta, \theta, \psi\} \in \mathbb{R}^{156}\), parameters for appearance \(\alpha \in \mathbb{R}^{50}\), spherical harmonics lighting \(l \in \mathbb{R}^{27}\), camera \(c \in \mathbb{R}^{3}\) (2D translation and isotropic scale of a weak-perspective camera), and style \(s \in \mathbb{R}^{512}\), generates an image of resolution \(P \times P\). Here, the FLAME parameters control all aspects related to the geometry of the face (i.e. width, height, expression, orientation, etc.), appearance and lighting parameters control the face color (i.e. skin tone, lighting, etc.), while the style vector \(s\) 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 the Flickr images (FFHQ) introduced in StyleGAN [32] and their corresponding FLAME parameters, appearance parameters, lighting parameters, and camera parameters. We obtain these parameters with a regression-based method derived from RingNet [53], trained with an additional photometric loss using a differentiable renderer [49]. In total, we use about 65,500 FFHQ images, paired with the corresponding parameters.

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. true generation process of our data is \(x \sim P(x|c_1, c_2)\) where \(P(x|c_1, c_2) = P(x|c_1) \cdot P(c_1) \cdot P(c_2)\). 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 cross-talk. Inductive bias introduced by condition representation in the context of conditional cross-talk is empirically evaluated in Section 5. This prompts us to find a ‘better’ representation of the conditioning variables.

**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 uses parts of the representation power of the generator to approximate classical rendering operations, which can be done learning-free and which are already part of several software packages (e.g. [40, 49]). The generator further needs to learn that rendering 3D models into images has few redundancies in its representation; namely, under a perspective projection,
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.

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 $s \in \mathbb{R}^{512}$ to model these factors. Note that original StyleGAN2 [33] has a similar vector $z$ with the same dimensionality, however, 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 a large inductive bias for this vector to capture background and appearance specific information, i.e. information that is not provided by our conditioning images.

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

Using the parametrization of a 3D model has several benefits, one of which, as mentioned before, allows for direct pixel-aligned supervision by representing the condition as mesh renderings. Besides that, it also allows us to impose further constraints on the generator output as detailed here.

We encourage GIF to inherit FLAME’s geometric disentanglement 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 $\alpha$ 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 and provides better disentanglement of the style vector from factors controlled by the FLAME parameters (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 $\alpha$, 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. Top and the 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 distribution 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 parameter control, it can directly be combined with existing FLAME-based application methods such as VOCA [18], 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 parameters to drive GIF for different appearance embeddings (see Figure 5). For more qualitative results and the full anima-
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 with an existing speech-driven facial animation method by generating face images for FLAME parameters obtained from VOCA [18]. Sample frames to highlight jaw pose variation. Please see the supplementary video for the full animation.

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. See the video for examples for both studies.

Ablation experiment: For the first 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 by using GIF, another by an ablated GIF model. Participants then select the generated image that corresponds best with the reference image. Table 1 shows that with 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 for detail on 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

| GIF | Vector cond. | No Texture interpolation | Normal rend. conditioning | Texture rend. conditioning |
|-----|--------------|-------------------------|---------------------------|---------------------------|
| 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.

| 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.

To see the supplementary video for the full animation.
randomly pair style vectors and FLAME parameters and conduct a perceptual study with AMT. Participants see a rendered FLAME image and a 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 \( s \) 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).

6. Discussion

GIF is based on the architecture of StyleGAN2 and hence inherits some of its limitations, like the requirement of the images to be roughly eye-centered. Although StyleGAN2 [34] has to some extent addressed this issue (among others) by removing the progressive training scheme, the data set of FFHQ images mostly has eyes in a specific image location. Hence, rotations of the head look eye-centered rotation, which involves a combination of 3D rotation and translation as opposed to pure rotation (neck centered). Adapting GIF for other than StyleGAN2 architecture or training it on a different data set to improve image quality is subject to future work.

As GIF uses the FLAME renderings for conditioning, these renderings must be similarly eye-centered as the FFHQ training data to be pixel-aligned with the generated image. For this, 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 and this often causes severe artifacts. Please refer to the supplementary material for some examples.

GIF requires a statistical 3D model (FLAME), and the possibility to associate the parametrization of this model to a large data set of high-quality images (FFHQ). While other classes of objects like human bodies [45] or animals [71] potentially also fulfill these requirements, it remains unclear how to apply GIF to general object categories.

GIF uses estimated model parameters from single images as supervision. Faulty associations sourcing from this regression-based approach potentially degrade the quality of the generated images and the disentanglement of different factors. 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 and hence models most skin color variations as part of the lighting rather than the appearance variation.

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.), resulting in jittery video sequences. However, our main goal is to have a FLAME-like model which associates semantically meaningful control to photo-realistic images. Training or refining GIF on temporal data with additional temporal constraints during training is subject to future work.

7. Conclusion

We present GIF, a high-quality generative 2D face model 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 high-quality 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 param-
eters and provides us with better temporal consistency when generating frames of FLAME sequences.

8. Acknowledgements

We thank H. Feng for preparing the training data, Y. Feng and S. Sanyal for support with the rendering and projection pipeline; and C. Khler, A. Chandrasekaran, M. Keller, M. Landry, C. Huang, A. Osman and D. Tzionas for fruitful discussions, advice and proofreading the manuscript. The work was partially supported by the International Max Planck Research School for Intelligent Systems (IMPRS-IS).

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 Meschcapade GmbH. PG has received funding from Amazon for using AWS GPUs. 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 International Conference on Automatic Face & Gesture Recognition (FG), pages 1–6, 2008. 2
[2] A. Bansal, S. Ma, D. Ramanan, and Y. Sheikh. RecycleGAN: Unsupervised video retargeting. In European Conference on Computer Vision (ECCV), pages 119–135, 2018. 3
[3] A. Bas and W. A. P. Smith. What does 2D geometric information really tell us about 3D face shape? International Journal of Computer Vision (IJCV), 2019. 5
[4] 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
[5] 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
[6] 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
[7] 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
[8] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In SIGGRAPH, pages 187–194, 1999. 2
[9] 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
[10] 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
[11] 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
[12] 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
[13] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Face-Warehouse: A 3D facial expression database for visual computing. IEEE Transactions on Visualization and Computer Graphics, 20(3):413–425, 2015. 2
[14] 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
[15] 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
[16] 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
[17] 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
[18] 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
[19] 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
[20] 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
[21] B. Egger. Semantic Morphable Models. PhD thesis, University of Basel, 2018. 5
[22] B. Egger, W. A. Smith, A. Tewari, S. Wuhrer, M. Zollhoefer, T. Beeler, F. Bernard, T. Bolkart, A. Kortylewski, S. Romdhan, et al. 3D morphable face models—past, present and future. Transaction on Graphics, 39(5), 2020. 2
[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. Bhattacharai, 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, 4

[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 StyleGAN. CoRR, abs/1912.04958, 2019. 2, 4, 5

[34] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of StyleGAN. 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. Baltrusaitis, 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. Aggarwal, J. Fursund, H. Li, R. Roberts, et al. pGAN: real-time avatars using dynamic textures. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia), 37(6):258–1, 2018. 2

[43] 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

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

[45] G. Pavlakos, V. Chouts, 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

[46] P. Paysan, R. Knothe, B. Amberger, 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

[47] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfelu, 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

[48] 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

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

[50] 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

[51] 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

[52] 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

[53] 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

[54] 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
[55] 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

[56] 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

[57] 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

[58] 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

[59] 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

[60] 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

[61] 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

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

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

[64] 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

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

[66] 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

[67] 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

[68] 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

[69] 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

[70] 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

[71] 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
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 \((\beta, \theta, \psi, \alpha)\), appearance \((\alpha)\) 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.

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 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.