Removing the Quality Tax in Controllable Face Generation

Yiwen Huang Zhiqiu Yu Xinjie Yi Yue Wang James Tompkin
Brown University

Abstract

3DMM conditioned face generation has gained traction due to its well-defined controllability; however, the trade-off is lower sample quality: Previous works such as DiscoFaceGAN and 3D-FM GAN show a significant FID gap compared to the unconditional StyleGAN, suggesting that there is a quality tax to pay for controllability. In this paper, we challenge the assumption that quality and controllability cannot coexist. To pinpoint the previous issues, we mathematically formalize the problem of 3DMM conditioned face generation. Then, we devise simple solutions to the problem under our proposed framework. This results in a new model that effectively removes the quality tax between 3DMM conditioned face GANs and the unconditional StyleGAN.
Project webpage: visual.cs.brown.edu/taxfreegan

1. Introduction

Face image generation has wide application in computer vision and graphics. Among different works in this area, deep learning generative model approaches are especially good at generating high-quality photo-realistic face images [17, 19]. However, generative models provide limited explicit control over their output due to their unsupervised nature, relying instead on latent space manipulation [21]. On the other hand, parametric models such as 3D Morphable Models (3DMMs) embed facial attributes in a disentangled parameter space, but their results lack photorealism [32].

In light of this, researchers have tried to build models that can synthesize high-resolution novel face images with control by combining 3DMM with generative modeling [1, 6, 8, 25, 39]. Existing attempts can be roughly divided into two categories: rigging and conditional generation. Rig-based methods attempt to align the 3DMM parameter space with the latent space of a pre-trained generative model [1, 39]. Sample quality is not compromised by controllability; however, controllability is limited by the completeness and disentanglement of the underlying latent space [43]. Conditional generation methods use 3DMM when training the generative model [6, 8, 25]. These offer improved controllability but reduced sample quality since additional constraints are imposed for 3DMM consistency and disentanglement.

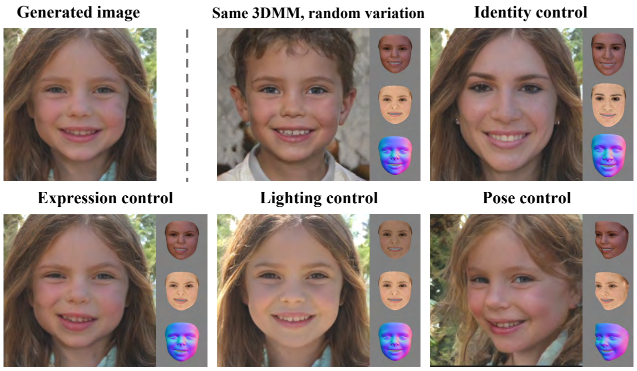

Figure 1. Past 3DMM-conditioned GANs show reduced image generation quality as a 'tax' for their added control. Our approach produces images of almost equivalent quality to unconditional generation while being at least as disentangled for control.

We investigate the family of 3DMM conditional GAN models. Deng *et al*. state that the quality drop in conditional models is an inevitable tax that we pay for controllability [6]. What causes this tax? We hypothesize that it is caused by overconstraint: that, to achieve consistency with the 3DMM conditioning *and* disentanglement among latent variables, current methods have unnecessary side effects that compromise quality. We challenge the claim of a 'quality tax' and show that it can be largely removed if the overconstraints can be identified and resolved. To this end, we formalize 3DMM conditioned face generation and identify minimal solutions that satisfy controllability and disentanglement.

Practically, we accomplish this with a differentiable 3DMM renderer [7] that by construction allows differentiable 3DMM parameter estimation from images. With this, we can directly minimize the mutual information between the distribution of 3DMM parameters and the distribution of images conditioned upon those 3DMM parameters. Once trained upon a StyleGAN2 base, this leads to a 3DMM-conditioned model that: (1) achieves significantly better FID scores than two SOTA methods (3.93 vs. 12.2), with a value that is almost equivalent to baseline unconditioned StyleGAN2 (3.78); and (2) also achieves equivalent or better disentanglement scores than two SOTA methods on two proposed metrics. Our findings effectively remove the quality tax of 3DMM-conditioned controllable face generation.

2. Related Work

Given the specificity of our contribution, we focus on a slice of works that explains 3DMM-conditioned GANs.

Face generation using GANs and disentangling. In 2017, with the introduction of PGGAN [16], the long-standing challenge of generating high resolution images had its first breakthrough, and Karras et al.'s StyleGAN family [17–19] has been the state-of-the-art in single domain image synthesis since then. A natural subsequent task is controlled generation of photorealistic images using StyleGAN. Despite numerous attempts [3, 22, 36, 37, 42–44], achieving both image quality and controllability remains challenging due to the lack of tractable semantics in StyleGAN's latent spaces.

3D prior for face modeling and synthesis. Numerous 3D methods for face generation exist. Among these, 3D Morphable Models (3DMMs) [11, 32] constitute a statistical approach that embeds human faces into a parameter space consisting of a set of principal components that represent factors including identity, expression, illumination, and pose. In contrast, Neural Radiance Fields (NeRFs) [28] generate photorealistic 3D scenes by leveraging a learned neural network to model the implicit 3D geometry and appearance of the target. While a number of 3D-aware models [4, 10] have incorporated NeRFs to synthesize facial images with pose variations, this approach is very computationally expensive.

3DMM-conditioned StyleGAN. The semantic interpretability of 3DMM offers the potential for controllable generation of faces. Proposed works combine 3DMM and StyleGAN to try to gain both semantic control and image quality. One type of such work [6, 25, 38] *conditions* their model training on the 3DMM parameter space, but the added control constraints the output quality. In contrast, another type of work [39] *rigs* the StyleGAN latent space, producing higher quality results but restricting control.

Disentangled representation learning (DRL). DRL [2] aims to represent and disentangle the constitutional factors lying in the data of interest. Many studies have sought to apply DRL to GAN for disentangled face synthesis. In particular, InfoGAN [5] and its variants [24, 31] attempt to maximize the mutual information between latent codes and generated samples to enforce disentangling. Peebles *et al.* [33] design a regularization term that encourages the Hessian of a model with respect to its input to be diagonal, thereby minimizing the interdependence of target factors.

3. Background and Problem Formulation

We define face images in a dataset $\hat{x} \in \mathcal{X}$. We also define a 3DMM code vector by $p = \{z_{\text{id}}, z_{\text{exp}}, z_{\text{illum}}, z_{\text{angle}}, z_{\text{trans}}\}$, a noise vector z, and a generator model $G(p,z) : \mathcal{P} \times \mathcal{Z} \rightarrow \mathcal{X}$.

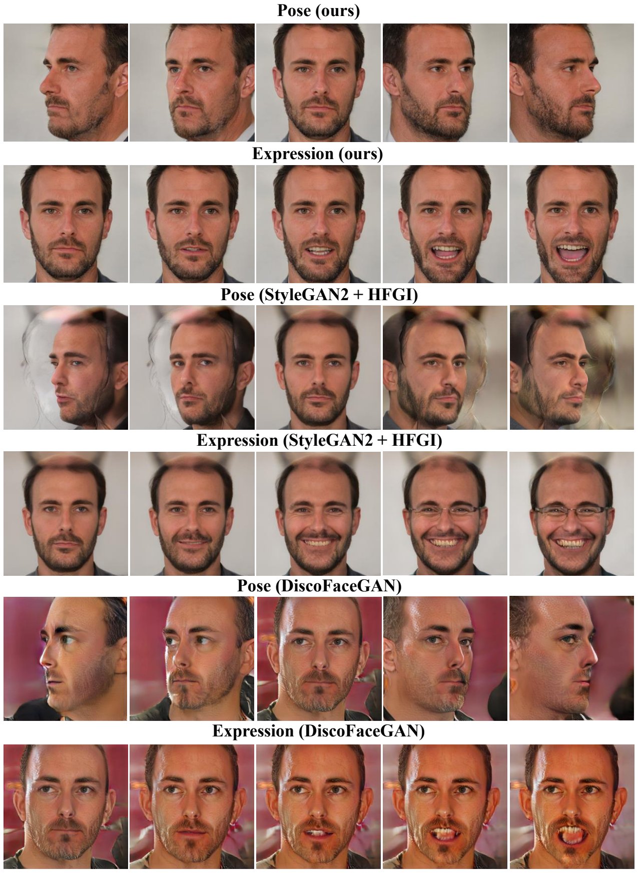

Figure 2. Extreme pose and expression variation with our model, StyleGAN2 using the SOTA post-hoc non-3DMM conditioning method HFGI [42], and DiscoFaceGAN [6]. DiscoFaceGAN has no inversion code to find style vector w for an image, so we provide a qualitative comparison: We use the same 3DMM parameters for the DiscoFaceGAN results as in ours, and vary the same controllers, but non-3DMM factors are determined by a random z vector.

The goal of conditional generation is to create photorealistic face images x according to p and z. Toward this goal, we concern ourselves with effective conditioning via 3DMM parameters p only; we leave open the disentangling of factors with no supervision. For our goal, we can form two related yet distinct objectives: *consistency* and *disentanglement*. But first, we explain why p is a difficult conditioning space.

3DMM representation. While p itself is an option for the consistency objective and conditioning G, previous studies show that the 3DMM parameter space $\mathcal{P}$ is suboptimal compared to a more image-based representation [8, 25]. Why is this? Given each component of $p = \{z_{\text{id}}, z_{\text{exp}}, z_{\text{illum}}, z_{\text{angle}}, z_{\text{trans}}\}$, z_{id} and z_{exp} determine the shape $\mathbf{S}$ and texture $\mathbf{T}$ of a face as follows [6]:

$$\mathbf{S} = \bar{\mathbf{S}} + \mathbf{B}_{\text{id}_s} z_{\text{id}_s} + \mathbf{B}_{\text{exp}} z_{\text{exp}} \quad \text{and} \quad \mathbf{T} = \bar{\mathbf{T}} + \mathbf{B}_{\text{id}_t} z_{\text{id}_t}$$

where $\bar{\mathbf{S}}$ and $\bar{\mathbf{T}}$ denote the average shape and texture, and $\mathbf{B}_{\text{id}_s}$, $\mathbf{B}_{\text{exp}}$, and $\mathbf{B}_{\text{id}_t}$ are the PCA bases of shape identity, facial expression, and texture. The definition of z_{id} and z_{exp}

depends on external factors such as $\mathbf{B}_{\text{id}_{s,t}}$ and $\mathbf{B}_{\text{exp}}$. Similarly, z_{illum} depends on the spherical harmonic basis **SH**. Without informing G of the external factors that each $z_i \in \mathcal{P}$ is defined upon, conditioning G directly on p imposes a challenge upon G to decipher the information encoded in p.

Tewari *et al*. noticed that using p as part of the optimization objective also leads to inferior results [39, 40]. They hypothesize that this is due to each $z_i \in \mathcal{P}$ having different perceptual effects in the image space. Since each z_i is defined w.r.t. different bases, the same magnitude of variation in different z_i might lead to different magnitudes of variation in image space. If we optimize the consistency object w.r.t. p directly, we gain consistency in $\mathcal{P}$ but not in image space.

Consistency. This objective requires that x is semantically consistent with p, *i.e*., p dictates the corresponding semantic factors in x. We follow the formulation in InfoGAN [5] and formalize the consistency objective as maximizing the mutual information $I(p;x)$ between p and x. This is defined as the difference between the entropy $H(p)$ and the conditional entropy $H(p\,|\,x)$:

$$\begin{aligned} I(p;x) &= H(p) - H(p\,|\,x) \\ &= \mathbb{E}_{x\sim G(p,z)}\left[\mathbb{E}_{p'\sim P(p|x)}[\log P(p'\,|\,x)]\right] + H(p) \\ &= \mathbb{E}_{p\sim P(p), x\sim G(p,z)}[\log P(p\,|\,x)] + H(p) \end{aligned} \tag{1}$$

The posterior $P(p|x)$ is not tractable in general GAN training, but Chen et al. show that $P(p|x)$ can be approximated by its variational lower bound [5]. As $H(p)$ does not depend on x, $H(p)$ is not optimizable and so is a constant.

For 3DMM conditioned face generation, the posterior becomes tractable when the generative distribution P_g becomes sufficiently close to the distribution of real face images. In such case, the posterior is exactly represented by a pretrained face reconstruction model [7] that can accurately predict p given x, allowing $I(p;x)$ to be directly optimized.

Past works propose proxy objectives instead of directly maximizing $I(p;x)$. These objectives maximize $I(p;x)$ up to some deterministic transformation on p. Deng *et al*. use imitative learning to enforces consistency on different components of p, using a combination of identity loss, landmark loss, spherical harmonic coefficient discrepancy for illumination, and skin color loss for albedo [6]. Further, Liu *et al*. proposed a consistency loss that minimizes the pixelwise difference between x and the image representation of p produced by a differentiable renderer [25].

We show that directly optimizing the mutual information objective is better than optimizing proxy objectives. Further, as the assumption that P_g is sufficiently close to the real image distribution does not hold in general early in training, we also introduce a progressive blending mechanism.

Disentanglement. Changing one semantic factor should not interfere with other semantic factors. Let $\mathcal{P} \cup \mathcal{Z} = \{z_0, z_1, \ldots, z_n\}$ where z_i denotes the latent code for an independent semantic factor. We formally define disentanglement following Peebles *et al*. [33]:

$$\frac{\partial^2 G}{\partial z_j \partial z_i} = 0 \quad \forall i \neq j \tag{2}$$

Suppose we define a subset of latent factors that control 3DMM factors; $z_i \in \mathcal{P}$. For these, disentanglement is achieved by construction via the consistency objective. The remaining problem is to disentangle unsupervised factors $z_j \in \mathcal{Z}$ from $z_i \in \mathcal{P}$. For example, 3DMM can control facial expression but not head hair; we must ensure that facial expression in p via z_i does not affect head hair length as controlled by z_j. Finally, as noted, the disentangling of unsupervised factors $z_j \in \mathcal{Z}$ from each other is an open question [26, 30] and does not relate to 3DMM conditioning.

In the simplest case where G is a scalar function and each semantic factor z_i is also a scalar, Eq. 2 indicates that the Hessian matrix $\mathbf{H}_G$ is diagonal. In such case, disentanglement can be directly encouraged by a Hessian penalty. A fast finite difference approximation of the penalty and a generalized version for vector-valued functions were also proposed [33]. However, it is observed that a Hessian penalty has a strong negative impact on image quality (measured by FID [13]) [33] and a solution to this problem is not yet clear.

As we found for consistency, disentanglement is also approximated by proxy objectives in previous work. Deng *et al*. proposed contrastive learning to approximate ${\partial^2 G}/{\partial z_i \partial z_{\text{exp}}} = 0 \forall i \neq \text{exp}$ and ${\partial^2 G}/{\partial z_{\text{hair,id}} \partial z_{\text{illum}}} = 0$ [6]. Liu *et al*. introduced disentangled training as an approximation of ${\partial^2 G}/{\partial z_{\text{id}} \partial z_i} = 0 \; \forall i \neq \text{id}$ [25]. We notice that all such approximations are restrictive; they degrade image quality and rely on hand-designed rules that only work for certain z_i, or attempt to encourage disentanglement through losses rather than through the construction of the network architecture.

To this end, we propose an alternative approach to the disentanglement problem. We neither attempt to directly penalize the non-diagonal entries of $\mathbf{H}_G$ [33] nor rely on proxy objectives to approximate a Hessian penalty [6, 25]. We show in the following section that, in practice, disentanglement can be achieved *for free* without any optimization via the inductive bias of a carefully designed network.

4. Method

4.1. Consistency via p Rendering & Estimation

We maximize Eq. 1 to enforce semantic consistency between p and x. However, there remains a design space of deterministic transformations on p to obtain a more amenable representation for conditioning and optimizing G. To this end, we use a differentiable renderer **RDR** [7] to derive a 3DMM representation that aligns with the image space perceptually, and is independent of external factors (Fig. 3).

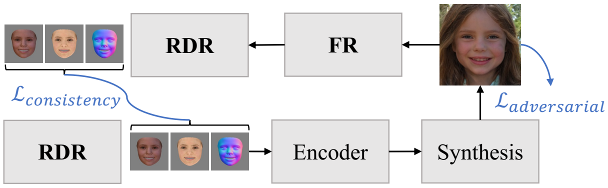

Figure 3. Our simple approach uses differentiable renderer **RDR** and 3DMM parameter estimator **FR** with only two objectives.

Specifically, we let **RDR** output the 3DMM rendered image r from p, the Lambertian albedo a, and the normal map n:

$$r,a,n = \mathbf{RDR}(p) \tag{3}$$

We define our 3DMM representation 'rep' as the Cartesian product of r, a and n: $\text{rep}(p) = r \times a \times n$. Given the new 3DMM representation, we update Eq. 1:

$$I(\text{rep}(p);x) = \mathbb{E}_{p\sim P(p),x\sim G(\text{rep}(p),z)}[\log P(\text{rep}(p)\,|\,x)] + C \tag{4}$$

where C is the constant term $H(\text{rep}(p))$.

Consistency loss. Given a pretrained face reconstruction model **FR** [7]:$\mathcal{X} \to \mathcal{P}$, we rewrite Eq. 4 as follows:

$$\mathcal{L}_{\text{consistency}} = \mathbb{E}_{p\sim P(p),x\sim G(\text{rep}(p),z)}\left[\|\text{rep}(\mathbf{FR}(x)) - \text{rep}(p)\|_{\text{p}}^{\text{p}}\right]. \tag{5}$$

The choice of p depends on our assumption about the functional form of the posterior. We follow common assumptions and assume Gaussian error, which leads to $\text{p}=2$. [9]

Liu *et al*. proposed an image-space consistency loss [25]:

$$\mathcal{L}_{\text{consistency}}^{\text{Liu}et al.} = \mathbb{E}_{p\sim P(p),x\sim G(r(p),z)}\left[\|x - r(p)\|_2^2\right]. \tag{6}$$

We show in our ablation study that this formulation of the consistency loss leads to significant quality degradation. Eq. 6 penalizes photorealism and encourages mode collapse: $\forall\, z$ given a fixed p, Eq. 6 pushes all x_z towards a single solution $r(p)$, therefore hindering the diversity of samples produced by G. Further, there is a domain gap between x and r as r is not photorealistic. A photorealistic face image often contains objects or phenomena (indirect illumination, eyeglasses, *etc*.) not modeled by the 3DMM. Eq. 6 is agnostic to such a domain gap and pushes x away from the real image distribution, thus compromising photorealism.

Our use of **FR** alleviates both these problems. **FR** essentially functions as a filter that removes all factors in x which are irrelevant to the 3DMM. As a result, these factors remain free variables and are not affected by our consistency loss.

Progressive blending. The posterior $P(p|x)$ can only be represented by **FR** when P_g is sufficiently close to the real image distribution. In early training with Eq. 5, x is not a realistic image and so $\mathbf{FR}(x)$ is nonsensical. This leads to instant collapse from ill-behaved $\mathcal{L}_{\text{consistency}}$ that is magnitudes larger than the adversarial loss, and from the consistency loss diverging in the first few training steps. To circumvent this problem, we introduce a progressive blending variant of Eq. 5, following the intuition that r is always a close enough approximation of the real face for **FR**:

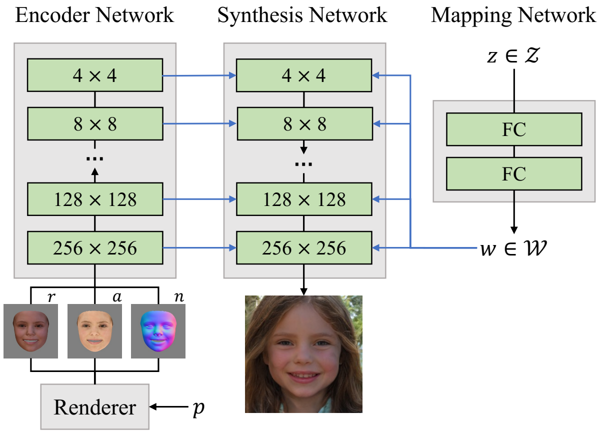

Figure 4. Our overall model architecture. We add an additional encoder network to condition the layer-wise synthesis process.

$$\mathcal{L}_{\text{consistency}}^{*} = \mathbb{E}_{p\sim P(p),x\sim G(\text{rep}(p),z)}[d] \tag{7}$$
$$d = \|\text{rep}(\mathbf{FR}(\alpha x + (1-\alpha)r(p))) - \text{rep}(p)\|_2^2$$

where α is a scalar that grows linearly from 0 to 1 in the first k training images. This initializes the input of **FR** to r, then the input gradually fades into x as the training progresses. We empirically find that this simple strategy is sufficient to solve the intractable posterior problem early in the training.

4.2. Structurally Disentangled Conditioning

Next, we discuss in detail how we use $\text{rep}(p)$ to condition G. We generate per-layer conditioning feature maps $c = \{c_1,\ldots,c_l\}$ using an encoder E, and inject each c_i into the corresponding layer of the synthesis network as an auxiliary input (Fig. 4). We show that our conditioning method approximates Eq. 2 without supervision [6,25], achieving disentanglement *for free* as an inductive bias of the network architecture.

Conditioning feature maps. We demonstrate our approach upon the common StyleGAN2 architecture [19]. We follow their design and split E into different resolution stages. For each resolution stage e_i of E, we produce two sets of feature maps c_{2i} and c_{2i+1} to condition the two synthesis layers of the corresponding resolution stage of the synthesis network:

$$e_i = \begin{cases} E_0(\text{rep}(p)) i=0 \\ E_i(e_{i-1}) i\neq 0 \end{cases} \tag{8}$$
$$c_{2i} = \text{toFeat}_{2i}(e_i)$$
$$c_{2i+1} = \text{toFeat}_{2i+1}(e_i)$$

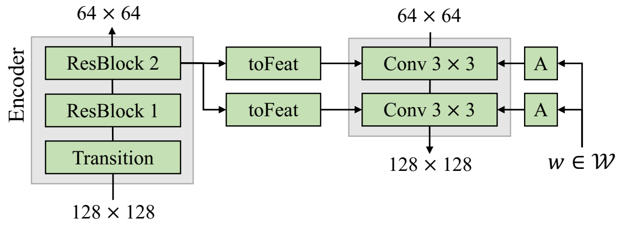

Figure 5. Encoder and feature injection module architecture.

We implement E_i as a sequence of a transition layer and two residual blocks (Fig. 5). 'toFeat' is implemented by a 1×1 convolution [23] with optional downsampling [18] and leaky ReLU activation [27]. See supplemental for more details.

Feature injection. We extend each synthesis layer l_i to take an auxiliary input c_{n-i} where n is the number of layers in the synthesis network. The synthesis layer in [19] is implemented by a stylized convolution where each channel f_j of the input feature maps f is scaled by s_{ij}. The per-layer scaling vector $s_i = \{s_{ij} \forall j\}$ is computed from the style vector w_i via an affine transformation. We note that the injected feature maps c_{n-i} need to be handled separately for stylization. This is because c_{n-i} is essentially an embedding of $\mathcal{P}$ while w_i is an embedding of $\mathcal{Z}$. It is clear that $\mathcal{P}$ is not controlled by $\mathcal{Z}$ and therefore c_{n-i} should not be subject to w_i. To this end, we simply fix the scaling of each channel of c_{n-i} to 1 for stylization.

In contrast to our feature injection-based conditioning, existing conditioning methods often involve manipulating the style vectors w^+. This can be done either by providing additional conditioning to the mapping network [6] or directly injecting conditioning to the $\mathcal{W}^+$ space [25]. Such style-based conditioning is problematic in two aspects:

1. There is no structural distinction between $\mathcal{P}$ and $\mathcal{Z}$ since both are encoded in $\mathcal{W}^+$. This necessitates additional disentanglement training objectives to decouple variation in $\mathcal{P}$ from variation in $\mathcal{Z}$. The disentanglement objectives are often ad hoc [6] and can compromise quality [6, 25].
2. The expressiveness of the $\mathcal{W}^+$ space is limited by its comparative low dimensionality. Encoding 'rep' in $\mathcal{W}^+$ requires high compression that might lead to information loss. We notice obvious discrepancies between x and r in previous work [25], and information loss might be a cause.

Our conditioning method avoids both problems and gives us disentanglement *for free*. In our method, c is a feature pyramid and each c_i has the same spatial dimensions as the input of the synthesis layer. Thus, we encode rep in c in high fidelity.

Disentanglement analysis. To simplify analysis, we omit various details from the StyleGAN2 [19] generator (weight demodulation, noise injection, equalized learning rate, *etc*.). We formulate each layer l_i of the synthesis network as:

$$l_i(p,z) = \mathbf{W}_i * [c_{n-i}(p); s_i(z) \odot \sigma(l_{i-1}(p,z))] + \mathbf{B}_i \tag{9}$$

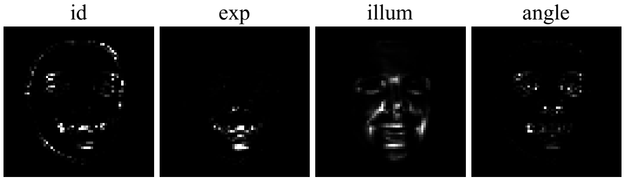

Figure 6. Finite difference approximation of the partial derivative of the injected 3DMM render features w.r.t. the 3DMM parameters $\partial c/\partial p$. Disentanglement is mostly successful: variation in c is shown by white regions, which are small and sparse.

$\mathbf{W}_i$ is the weight tensor of l_i, $\mathbf{B}_i$ is the bias tensor of l_i, $*$ denotes convolution, $\odot$ denotes the Hadamard product, and σ is the activation function. There are two terms in l_i that depend on p: c_{n-i} and $\sigma(l_{i-1})$. First, we analyze disentanglement w.r.t. c_{n-i}:

$$\begin{aligned}
\frac{\partial^2 l_i}{\partial_z \partial c_{n-i}} &= \frac{\partial}{\partial_z}\left(\frac{\partial}{\partial c_{n-i}}(\mathbf{W}_i * [c_{n-i}; s_i \odot \sigma(l_{i-1})] + \mathbf{B}_i)\right) \\
&= \frac{\partial}{\partial_z}\left(\mathbf{W}_i * \frac{\partial}{\partial c_{n-i}}[c_{n-i}; s_i \odot \sigma(l_{i-1})]\right) \\
&= \frac{\partial}{\partial_z}(\mathbf{W}_i * [I; 0]) \\
&= 0
\end{aligned} \tag{10}$$

We see that variation in c_{n-i} is perfectly disentangled from variation in z, therefore any non-zero $\frac{\partial^2 l_i}{\partial_z \partial_p}$ must be the result of variation in $\sigma(l_{i-1})$:

$$\begin{aligned}
\frac{\partial^2 l_i}{\partial z \partial p} &= \frac{\partial^2 l_i}{\partial z \partial \sigma(l_{i-1})} \frac{\partial \sigma(l_{i-1})}{\partial p} \\
&= \left(\mathbf{W}_i * \left[0; \frac{\partial s_i}{\partial z}\right]\right) \frac{\partial \sigma(l_{i-1})}{\partial p}
\end{aligned} \tag{11}$$

We examine the behavior of variation in p:

$$\begin{aligned}
\frac{\partial \sigma(l_{i-1})}{\partial p} &= \frac{\partial \sigma(l_{i-1})}{\partial l_{i-1}} \frac{\partial l_{i-1}}{\partial p} \\
&= \frac{\partial \sigma(l_{i-1})}{\partial l_{i-1}} \left(\mathbf{W}_{i-1} * \left[\frac{\partial c_{n-i+1}}{\partial p}; s_{i-1} \odot \frac{\partial \sigma(l_{i-2})}{\partial p}\right]\right)
\end{aligned} \tag{12}$$

This analysis on $\frac{\partial \sigma(l_{i-1})}{\partial p}$ applies recursively to $\frac{\partial \sigma(l_{i-2})}{\partial p}$; thus, $\frac{\partial^2 G}{\partial z \partial p} \to 0$ if $\forall i. \frac{\partial c_i}{\partial p} \to 0$.

In practice, we empirically find that small variation in p does lead to little total variation in c. Variation in c tends to be highly localized to small affected regions dictated by p, with little variation otherwise (Fig. 6). This is likely the combination effect of localized variation in rep w.r.t. p and the inductive bias of locality of a convolutional encoder. We do not consider $\frac{\partial^2 G}{\partial p \partial z}$ as disentanglement in this direction is automatically enforced by $\mathcal{L}_{\text{consistency}}$ when pairing each p with a set of different zs.

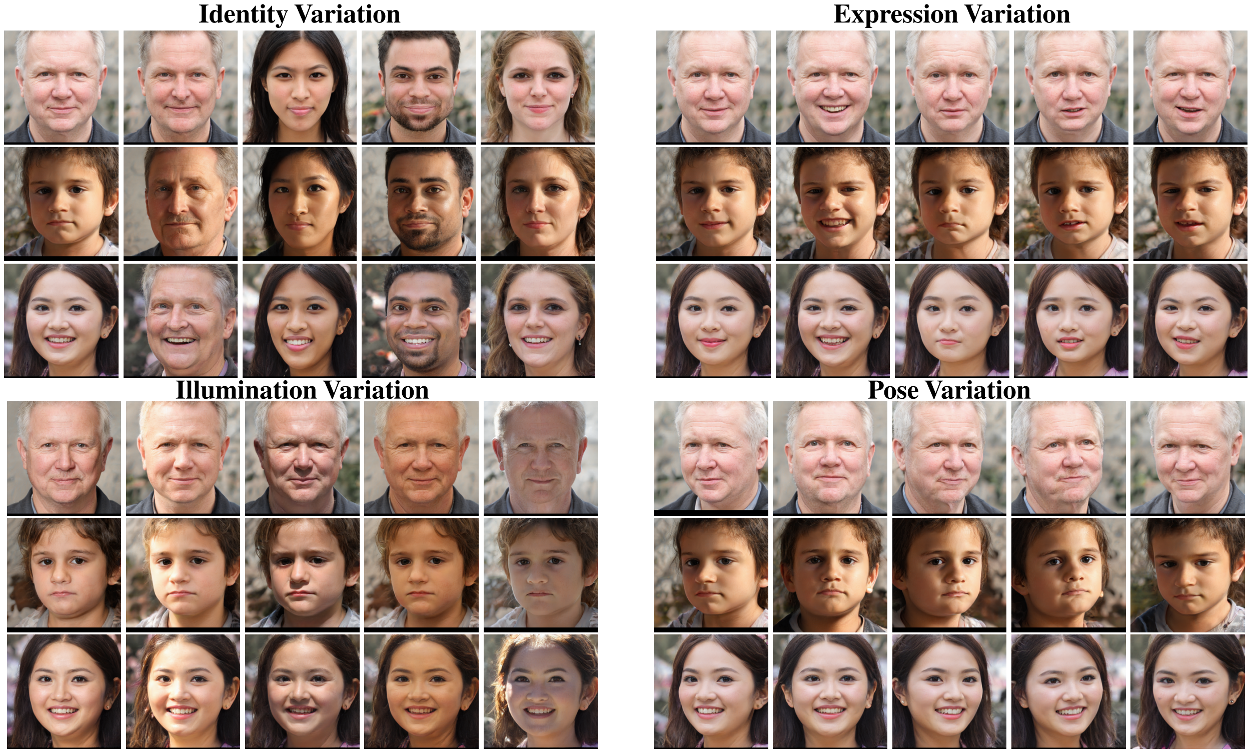

Figure 7. Generated face samples with control as output from our model. To attempt to reduce any impression of cherry picking, we use *the same* three input faces for each 3DMM attribute across five edit (columns). While some unwanted variation remains, identity, expression, illumination, and angle are controlled with high fidelity and no apparent visual artifacts.

5. Experiments

Data and Face Reconstruction We use FFHQ [18] at 256 × 256 resolution to generate our training data. To preprocess the data, we follow the method of Deng et al. [7]: We detect facial landmarks in all FFHQ images with MTCNN [45] and perform face alignment based on the landmarks detected. We derive 3DMM coefficients for each image. In the training stage, we use the aligned images as inputs and the corresponding 3DMM coefficients as training labels. Following DiscoFaceGAN [6], we use the pretrained face reconstruction model from Deng *et al*. [7] as **FR**.

Baselines We compare model performance against baselines in terms of generation quality and semantic disentanglement for editing. We use StyleGAN2 and two state-of-the-art 3DMM-based generative models, DiscoFaceGAN (DFG) [6] and 3D-FM GAN [25], along with other frontalization methods [12, 15, 34, 41, 46]. As the leading SOTA method 3D-FM GAN does not have public code or models, comparison is difficult. Where possible, we took results from their paper, but some quantitative metrics could only be computed for our model and for DiscoFaceGAN. We do not compare against rigging-based methods like StyleRig [39] as their controllability is upperbounded by the disentanglement and completeness of the existing StyleGAN2 latent space.

Controlled generation Our model achieves highly controllable generation while preserving StyleGAN's ability to generate highly photorealistic images (Fig. 7). We can see that our model can produce photorealistic faces with diverse races, genders, and ages and control over each of the 3DMM attributes. Particularly, we use the same three people for all attribute edits; this shows that our model can perform robust generation with high quality. Fig. 8 compares the images generated by our model conditioned on the same p but different z. The identity, expression, pose, and illumination are preserved while all other attributes are modified. This shows that there is little overlap between attributes controlled by p and z; that our model gains control over target attributes.

Real image inversion and editing Following [25], we test our model's ability to embed real images into its latent space and perform disentangled editing (Fig. 9). On zooming, we see that our model produces the sharpest images and that they align closely with the target references from the 3DMM renderer. While DiscoFaceGAN completely collapses on this input, 3D-FM GAN gives blurry and sometimes non-photorealistic outputs (e.g., under pose change).

We also compare on the task of face frontalization by simply rotating the 3DMM camera to identity (Fig. 10). Our model significantly improves frontalized quality against most methods, and compared with the state-of-the-art face manipulation models [6, 25], our model produces better identity-preserved faces in a more precise frontal view.

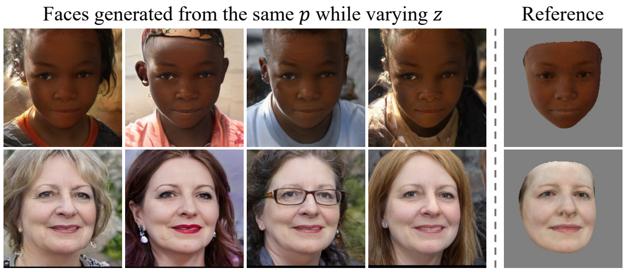

Figure 8. Resampling the noise vector z with the same set of 3DMM coefficients p shows high facial consistency, while other unsupervised factors like hair, hat, eyeglasses, and background vary with z.

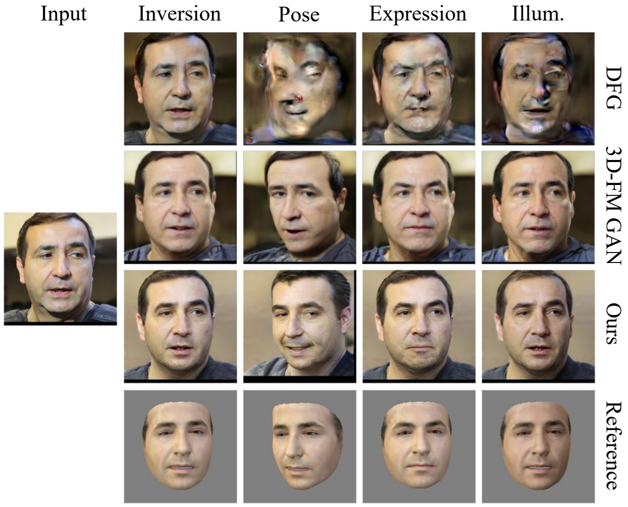

Figure 9. Real face editing. After inversion, for which all methods show some error, our model shows competitive ability. In both inversion and individual attribute edits, our model generates images that are faithful to the 3DMM renderer outputs and the input.

5.1. Quantitative Comparison

We evaluate the performance of our model in terms of quality and disentanglement. For image quality, we compute the Fréchet inception distance (FID) [13] and Precision and Recall (P&R) [20, 35] against the entire FFHQ dataset as a measure of the generation quality. Our model outperforms the two state-of-the-art baselines, yielding an FID much closer to the original StyleGAN trained on 256 × 256 FFHQ dataset (Tab. 1). Precision and recall indicate that our model has achieved near-StyleGAN-level image generation results while controlling and disentangling facial attributes.

Disentanglement Score Introduced in DiscoFaceGAN, this quantifies the disentanglement efficacy of each of the four 3DMM-controlled identity, expression, illumination, and angle attributes. Due to ambiguity in the derivation of this score [6], please see supplemental for details.

Table 1. Our conditioning provides control and almost equivalent quality to unconditioned baseline StyleGAN2. Two baseline 3DMM conditioning approaches do not produce comparable quality in terms of FID. P&R were introduced after StyleGAN1 and thus these numbers are missing from DiscoFaceGAN (built on StyleGAN1).

Method	FID↓	Precision↑	Recall↑
StyleGAN2	3.78	0.692	0.431
Ours	3.93	0.549	0.531
DiscoFaceGAN	12.9	-	-
3D-FM GAN	12.2	-	-

For attribute vector $u_i \in \{z_{\text{id}}, z_{\text{exp}}, z_{\text{illum}}, z_{\text{angle}}\}$, we first randomly sample 1K sets of the other three attribute vectors, denoted by $u_{\{j\}} = \{u_j : j = 1,...,4, j \neq i\}$. Then, for each set of $u_{\{j\}}$, we randomly sample 10 u_i. In total, we have 10K 3DMM coefficients and hence generate 10K images. Then, we re-estimate u_i and $u_{\{j\}}$ using the 3D reconstruction network [7]. For each attribute, we compute the L2 norm of the difference between each u and the mean u vector and get the mean L2 norm in each of the 1K sets. We then get σ_{u_i} and σ_{u_j}'s by averaging the corresponding mean L2 norm over the 1K sets and normalize them by the L2 norm of the mean u vector computed on the entire FFHQ dataset. Finally, we compute the disentanglement score:

$$DS(u_i) = \prod_{j, j \neq i} \frac{\sigma_{u_i}}{\sigma_{u_j}} \tag{13}$$

A high DS indicates that when an attribute vector is modified, only the corresponding attribute is changed on the generated image while all other attributes remain unchanged. Our model outperforms DiscoFaceGAN by large margins in identity, expression, and pose (angle) control (Table 2).

DCI This metric was introduced in StyleSpace [43]. Given a set of attributes and a latent space, *disentanglement* measures the extent to which each latent dimension controls at most one attribute, *completeness* measures the extent to which each attribute is controlled by at most one latent dimension, and *informativeness* measures how well attributes can be correctly predicted from a given latent representation.

To calculate DCI, we first sample 35K 3DMM coefficient vectors from FFHQ and generate corresponding images using these vectors. Then, we annotate the images by 8 binary classifiers trained on CelebA [18] that can be controlled by 3DMM coefficients, and train a gradient boosting classifier to predict the 3DMM coefficient vectors from the annotations.

Our model outperforms DiscoFaceGAN (Table 3). DCI indicates that our model establishes a better one-to-one relationship between the attributes and 3DMM coefficients, leading to a more disentangled 3DMM parameter space.

5.2. Ablation Study

We modify our untouched model (denoted **Config-A**) in three different ways to investigate its performance. All ablations are conducted on the 128 × 128 version of FFHQ [18], and all ablation models are trained on 5M real images.

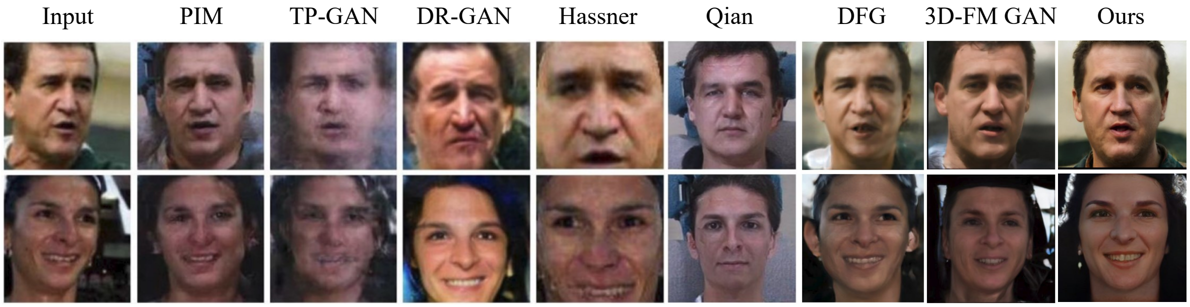

Figure 10. Face frontalization comparisons with DiscoFaceGAN (DFG), 3D-FM GAN and other models on LFW images [14]. Our model achieves a good balance between image fidelity and frontal pose positions.

Table 2. Disentanglement Score comparisons with DiscoFaceGAN across four 3DMM-controlled attributes.

Method	$DS_{\text{id}}\uparrow$	$DS_{\text{exp}}\uparrow$	$DS_{\text{illum}}\uparrow$	$DS_{\text{angle}}\uparrow$
DiscoFaceGAN	0.37	1.64	47.9	829
Ours	**1.02**	**3.22**	**48.7**	**1245**

Table 3. DCI metric comparisons.

Method	Disentanglement↑	Completeness↑	Informativeness↑
DiscoFaceGAN	0.66	0.73	0.98
Ours	**0.83**	**0.78**	**0.99**

Table 4. Ablation FID and Disentanglement Score comparisons.

Method	Quality	Disentanglement Score			
	FID↓	id↑	exp↑	illum↑	angle↑
Config-A	**8.73**	1.07	3.19	49.5	1402
Config-B	17.5	**1.14**	**5.97**	**57.2**	**1964**
Config-C	10.9	0.694	2.71	29.2	1142
Config-D	13.3	0.41	1.25	23.2	690

Config-B: Conditional discriminator. The 3DMM condition p or $\text{rep}(p)$ can be used to condition the discriminator D similarly to G. However, all past works [6, 8, 25] do not condition D; it is unclear whether this is an intentional design choice. To our surprise, conditioning D leads to significantly worse FID [13], contradicting the common belief that conditioning is always beneficial [29]. We experiment with various conditioning methods, all of which degrade FID considerably. This might be the result that the conditional distribution is undersampled. Unlike traditional class conditional generation where thousands of samples are available for a single condition, we essentially have one real sample for each p. The scarcity of samples might outweigh the benefit of extra condition information. Nevertheless, this config has improved disentanglement performance.

Config-C: Alternative consistency loss. We swap our consistency loss with Eq. 6 proposed by Liu *et al*. [25]. As expected, this change leads to inferior FID. Our model converges faster early in the training using this alternative consistency loss, but the FID quickly plateaus and is later surpassed by Config-A. The initial quick convergence is likely due to the lack of progressive blending, which results in a stronger learning signal early on. However, the overconstrained nature of Eq. 6 eventually impedes the model from further improving.

Config-D: One-layer feature injection. We remove feature injection from all synthesis layers except the layer in the 4×4 stage. This allows our model to emulate the behavior of a traditional conditional generator [19, 29]. We observe drastic performance drop in disentanglement compared to Config-A, indicating that our per-layer feature injection is crucial to disentanglement. Interestingly, we also observe a degradation in FID and poor adherence to p early in the training. Without per-layer injection, G has to rely exclusively on the global features c_{n-1} that we inject to the first synthesis layer, and any subtle variation in c_{n-1} will be amplified by each layer afterwards, resulting in poor disentanglement. The degradation in FID is likely due to the lack of a feature pyramid, that some of the network capacity of the synthesis network is wasted on decoding the highly compressed c_{n-1}.

6. Conclusion and Future Work

We present a simple conditional model derived from a mathematical framework for 3DMM conditioned face generation. Our model shows strong performance in both quality and controllability, reducing the need to choose between the two and making control 'tax free'. Furthermore, our mathematical framework can be applied to future explorations in conditional generation, allowing future investigators to analyze other 3DMMs rigorously. However, our model does not come without limitations. Unlike 3D-FM GAN [25], our model is not specifically designed for image editing. Thus, faces suffer the same inversion accuracy *vs*. editability tradeoff as StyleGAN [17–19]. Future work might consider applying the image editing techniques proposed by Liu *et al*. [25] to our model for better face editing.

Acknowledgements The authors thank Yu Deng for generous correspondence and assistance in correctly reimplementing Disentanglement Score metrics. Funding was provided by a Brown Office of the Vice President for Research Seed award.

References

[1] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (ToG)*, 40(3):1–21, 2021. 1

[2] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013. 2

[3] Amit H Bermano, Rinon Gal, Yuval Alaluf, Ron Mokady, Yotam Nitzan, Omer Tov, Oren Patashnik, and Daniel Cohen-Or. State-of-the-art in the architecture, methods and applications of stylegan. In *Computer Graphics Forum*, volume 41, pages 591–611. Wiley Online Library, 2022. 2

[4] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5799–5809, 2021. 2

[5] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *CoRR*, abs/1606.03657, 2016. 2, 3

[6] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[7] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 1, 3, 4, 6, 7

[8] Partha Ghosh, Pravir Singh Gupta, Roy Uziel, Anurag Ranjan, Michael J Black, and Timo Bolkart. Gif: Generative interpretable faces. In *2020 International Conference on 3D Vision (3DV)*, pages 868–878. IEEE, 2020. 1, 2, 8

[9] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org. 4

[10] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985*, 2021. 2

[11] Yudong Guo, Jianfei Cai, Boyi Jiang, Jianmin Zheng, et al. Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. *IEEE transactions on pattern analysis and machine intelligence*, 41(6):1294–1307, 2018. 2

[12] Tal Hassner, Shai Harel, Eran Paz, and Roee Enbar. Effective face frontalization in unconstrained images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4295–4304, 2015. 6

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 3, 7, 8

[14] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008. 8

[15] Rui Huang, Shu Zhang, Tianyu Li, and Ran He. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *Proceedings of the IEEE international conference on computer vision*, pages 2439–2448, 2017. 6

[16] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017. 2

[17] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021. 1, 2, 8

[18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2, 5, 6, 7, 8

[19] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 1, 2, 4, 5, 8

[20] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems*, 32, 2019. 7

[21] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. Fader networks: Manipulating images by sliding attributes. *CoRR*, abs/1706.00409, 2017. 1

[22] Eric-Tuan Le, Edward Bartrum, and Iasonas Kokkinos. Stylemorph: Disentangled 3d-aware image synthesis with a 3d morphable styleGAN. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[23] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 5

[24] Zinan Lin, Kiran Thekumparampil, Giulia Fanti, and Sewoong Oh. Infogan-cr and modelcentrality: Self-supervised model training and selection for disentangling gans. In *international conference on machine learning*, pages 6127–6139. PMLR, 2020. 2

[25] Yuchen Liu, Zhixin Shu, Yijun Li, Zhe Lin, Richard Zhang, and SY Kung. 3d-fm gan: Towards 3d-controllable face manipulation. In *European Conference on Computer Vision*, pages 107–125. Springer, 2022. 1, 2, 3, 4, 5, 6, 8

[26] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019. 3

[27] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. 5

[28] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[29] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 8

[30] Weili Nie, Tero Karras, Animesh Garg, Shoubhik Debnath, Anjul Patney, Ankit B Patel, and Anima Anandkumar. Semi-supervised stylegan for disentanglement learning. In *Proceedings of the 37th International Conference on Machine Learning*, pages 7360–7369, 2020. 3

[31] Jonasz Pamuła. Progressive training of gans with mutual information penalty. 2

[32] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pages 296–301. Ieee, 2009. 1, 2

[33] William Peebles, John Peebles, Jun-Yan Zhu, Alexei A. Efros, and Antonio Torralba. The hessian penalty: A weak prior for unsupervised disentanglement. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020. 2, 3

[34] Yichen Qian, Weihong Deng, and Jiani Hu. Unsupervised face normalization with extreme pose and expression in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9851–9858, 2019. 6

[35] Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31, 2018. 7

[36] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE transactions on pattern analysis and machine intelligence*, 44(4):2004–2018, 2020. 2

[37] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1532–1540, June 2021. 2

[38] Keqiang Sun, Shangzhe Wu, Zhaoyang Huang, Ning Zhang, Quan Wang, and HongSheng Li. Controllable 3d face synthesis with conditional generative occupancy fields. *arXiv preprint arXiv:2206.08361*, 2022. 2

[39] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. *CoRR*, abs/2004.00121, 2020. 1, 2, 3, 6

[40] Ayush Tewari, Mohamed Elgharib, Mallikarjun B R., Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. Pie: Portrait image embedding for semantic control, 2020. 3

[41] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1415–1424, 2017. 6

[42] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity gan inversion for image attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11379–11388, 2022. 2

[43] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12863–12872, June 2021. 1, 2, 7

[44] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3121–3138, 2022. 2

[45] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multi-task cascaded convolutional networks. *CoRR*, abs/1604.02878, 2016. 6

[46] Jian Zhao, Yu Cheng, Yan Xu, Lin Xiong, Jianshu Li, Fang Zhao, Karlekar Jayashree, Sugiri Pranata, Shengmei Shen, Junliang Xing, et al. Towards pose invariant face recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2207–2216, 2018. 6