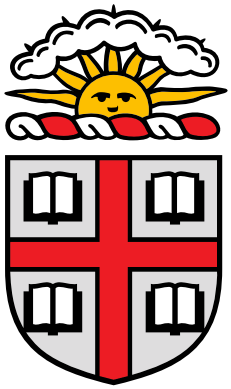
Quality or Control: Why Not Both? Rethinking 3DMM-Conditional Face Generation

---

Author
Xinjie Yi

Advisor
James Tompkin
Reader
Daniel Ritchie

A Thesis submitted in partial fulfillment of the requirements for Honors in the Department of Computer Science at Brown University



Brown University
Providence, Rhode Island

May 2023

Author:
Xinjie Yi

Advisor:
James Tompkin
Reader:
Daniel Ritchie

Institute:
Brown University, Providence, Rhode Island

ABSTRACT

3DMM-conditioned face generation has gained traction due to its well-defined controllability; however, this comes at a cost of lower sample quality. Previous works such as DiscoFaceGAN and 3D-FM GAN have showed a significant FID gap when compared to the unconditional StyleGAN, suggesting a trade-off between quality and controllability. This thesis challenges the conventional wisdom and proposes a new model that effectively removes the quality tax between 3DMM-conditioned face GANs and the unconditional StyleGAN. We mathematically formalize the issue of 3DMM-conditioned face generation to pinpoint previous challenges, and devise simple solutions within our proposed framework. The results demonstrate that quality and controllability can coexist, opening new possibilities for 3DMM-conditioned face generation.

ACKNOWLEDGMENTS

I am deeply grateful to my thesis advisor, Professor James Tompkin, for providing invaluable guidance, patience, and support throughout my research. His expertise and mentorship have been instrumental in helping me achieve my research goals.

I would like to thank my collaborators, Yiwen Huang, Zhiqiu Yu, and Yue Wang, for their significant contributions to this research. Their hard work and company have made this journey a truly rewarding one.

I would also like to express my gratitude towards Professor Daniel Ritchie for serving as my thesis reader. His insightful feedback has further improved the quality of this thesis.

A special thank goes to Professor Stuart Geman, who has played an integral role in my undergraduate research experience. His wisdom and encouragement have helped me grow both as a researcher and as an individual.

Finally, I would like to thank my family and friends for their unwavering love and support throughout. Their encouragement and belief in me have constantly been my source of inspiration and motivation.

CONTENTS

1 INTRODUCTION

The field of computer vision and graphics has witnessed a surge of interest in face image generation owing to its broad scope of applications. Among different works in this area, neural generative modeling approaches, notably generative adversarial networks (GANs) [1], have proven especially effective in generating high-quality, photorealistic face images [2, 3]. However, these models provide limited explicit control over their output, relying instead on latent space manipulation due to their unsupervised nature [4]. 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 [5].

In light of this, researchers have attempted to build models that can synthesize high-resolution novel face images with control by combining 3DMM with generative modeling [6–10]. Existing approaches can be roughly divided into two categories: rigging and conditional generation. Rig-based methods align the 3DMM parameter space with the latent space of a pre-trained generative model [8, 9]. These methods maintain high sample quality but limit controllability due to the completeness and disentanglement of the latent space [11]. In contrast, conditional generation methods use the 3DMM when training the generative model [6, 7, 10]. They offer improved controllability yet compromised sample quality since additional constraints are imposed upon the generated samples for 3DMM consistency and disentanglement.

This thesis aims to investigate the family of 3DMM-conditional GANs. 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 is, 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 an inherent "quality tax" and show that it can be largely eliminated if the overconstraints can be identified and resolved. To this end, we formalize 3DMM-conditioned face generation and identify minimal solutions that satisfy both controllability and disentanglement.

In practice, we implement a differentiable 3DMM renderer [12], which by construction enables differentiable 3DMM parameter estimation from images. Using this, we can directly minimize the mutual information between the distribution of 3DMM parameters and the distribution of images conditioned upon these parameters. Once trained on a StyleGAN2 base, this results in a 3DMM-conditioned model that: (1) achieves significantly better FID scores than two state-of-the-art methods (4.51 vs. 12.2), nearly matching the unconditioned StyleGAN2 baseline (3.78); and (2) obtains comparable or superior disen-

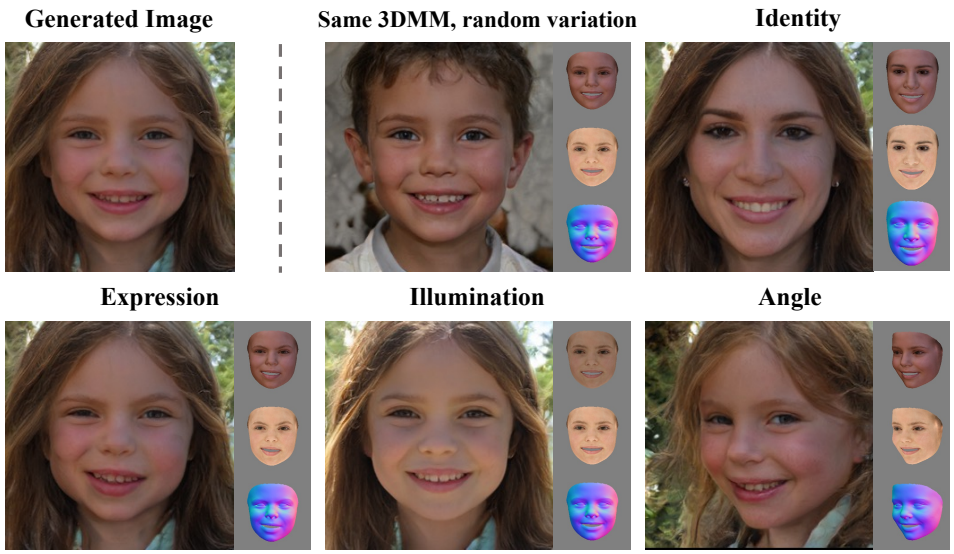


Figure 1.1: 3DMM-conditioned GANs show reduced image generation quality as a "tax" for their added control. This tax is not inevitable. Our approach produces images of almost equivalent quality to unconditional generation while being at least as disentangled for control.

tanglement scores to two state-of-the-art methods on two proposed metrics. Our findings effectively eliminate the quality tax of 3DMM-conditioned GAN models.

Our key contributions are threefold:

- We propose a mathematical framework for 3DMM-conditioned face generation, unifying existing methods within this formulation. This allows us to rigorously analyze the consistency and the disentanglement behavior.
- We derive novel methods for achieving consistency and disentanglement from our mathematical framework. We evaluate the limitations of existing methods, and demonstrate that our approaches are both theoretically sound and practically favorable when compared to previous work.
- We showcase a StyleGAN2-based model, trained by our methods, that achieves state-of-the-art FID while preserving the full controllability of 3DMM.

This thesis is a collaborative effort of co-authors Yiwen Huang, Zhiqiu Yu, Yue Wang, and James Tompkin. It is mainly adapted from our conference paper submission for ICCV 2023.

2

RELATED WORK

In this section, we discuss a relevant slice of work that explains 3DMM-conditioned GANs, given the specificity of our contributions.

2.1 GANS AND DISENTANGLING FOR FACE GENERATION

Generative Adversarial Networks (GANs) [1] are widely used to generate photorealistic images. In 2017, the introduction of PGGAN [13] marked a significant milestone in generating high-resolution images. Since then, Karras et al.'s StyleGAN family [2, 3, 14] has been the state-of-the-art in single-domain image synthesis. One natural subsequent task in this field is controlled generation of photorealistic images using StyleGAN. However, achieving both high image quality and controllability remains challenging, primarily due to the lack of tractable semantics in StyleGAN's latent spaces. Despite numerous attempts [11, 15–17], this issue remains unresolved.

2.2 3D PRIOR FOR FACE MODELING AND GENERATION

Numerous 3D methods for face generation exist. Among these, 3D Morphable Models (3DMMs) [5, 18] 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) [19] 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 [20, 21] have incorporated NeRFs to synthesize facial images with pose variations, this approach is very computationally expensive.

2.3 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, 7, 22] *conditions* their model training on the 3DMM parameter space, but the added control constraints the output quality. In contrast, another type of work [8] *rigs* the StyleGAN latent space, producing higher quality results but restricting control.

2.4 DISENTANGLED REPRESENTATION LEARNING

Disentangled representation learning (DRL) [23] 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 [24] and its variants [25, 26] attempt to maximize the mutual information between latent codes and generated samples to enforce disentangling. Peebles et al. [27] 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

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 $G(p, z) : \mathcal{P} \times \mathcal{Z} \to \mathcal{X}$. 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.

3.1 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 [7, 10]. 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]:

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

where $\overline{\mathbf{S}}$ and $\overline{\mathbf{T}}$ 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 $\mathbf{SH}$. 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 [8, 28]. 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 the image space.

3.2 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 [24] and formalize the consistency objective as maximizing the mutual information $I(p; x)$ between p and x:

$$\begin{aligned} I(p; x) &= H(p) - H(p \mid x) \\ &= \mathbb{E}_{x \sim G(p,z)} \left[\mathbb{E}_{p' \sim P(p|x)} \left[\log P(p' \mid x) \right] \right] + H(p) \\ &= \mathbb{E}_{p \sim P(p), x \sim G(p,z)} \left[\log P(p \mid x) \right] + H(p) \end{aligned} \tag{3.1}$$

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 [24]. 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 [12] 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 [7].

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.

3.3 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. [27]:

$$\frac{\partial^2 G}{\partial z_j \partial z_i} = 0 \qquad \forall\, i \neq j \tag{3.2}$$

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 disen-

tangling of unsupervised factors $z_j \in \mathcal{Z}$ from each other is an open question [29, 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. 3.2 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 [27]. However, it is observed that a Hessian penalty has a strong negative impact on image quality (measured by FID [31]) [27] 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 $\partial^2 G / \partial z_{\text{id}} \partial z_i = 0 \ \forall \, i \neq \text{id}$ [7]. 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 $\mathbf{H}_G$ [27] nor rely on proxy objectives to approximate a Hessian penalty [6, 7]. 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. 3.1 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** [12] to derive a 3DMM representation that aligns with the image space perceptually, and is independent of external factors. Specifically, we let **RDR** output the 3DMM rendered image r from p, the Lambertian albedo a, and the normal map n:

$$r, a, n = \mathbf{RDR}(p) \tag{4.1}$$

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. 3.1:

$$I(\text{rep}(p); x) = \mathbb{E}_{p\sim P(p), x\sim G(\text{rep}(p), z)} \left[\log P(\text{rep}(p) \mid x)\right] + C \tag{4.2}$$

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

4.1.1 *Consistency Loss*

Given a pretrained face reconstruction model **FR** [12]: $\mathcal{X} \to \mathcal{P}$, we rewrite Eq. 4.2 as follows:

$$\mathcal{L}_{\text{consistency}} = \mathbb{E}_{p\sim P(p), x\sim G(\text{rep}(p), z)} \left[\|\text{rep}(\mathbf{FR}(x)) - \text{rep}(p)\|_{\text{p}}^{\text{p}}\right] \tag{4.3}$$

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 $\text{p} = 2$ [32].

Liu et al. proposed an image-space consistency loss [7]:

$$\mathcal{L}_{\text{consistency}}^{\text{Liu } et\ al.} = \mathbb{E}_{p\sim P(p), x\sim G(r(p), z)} \left[\|x - r(p)\|_2^2\right]. \tag{4.4}$$

We show in our ablation study that this formulation of the consistency loss leads to significant quality degradation. Eq. 4.4 penalizes photorealism and encourages mode collapse: $\forall\, z$ given a fixed p, Eq. 4.4 pushes all x_z towards a single solution $r(p)$, therefore hinder-

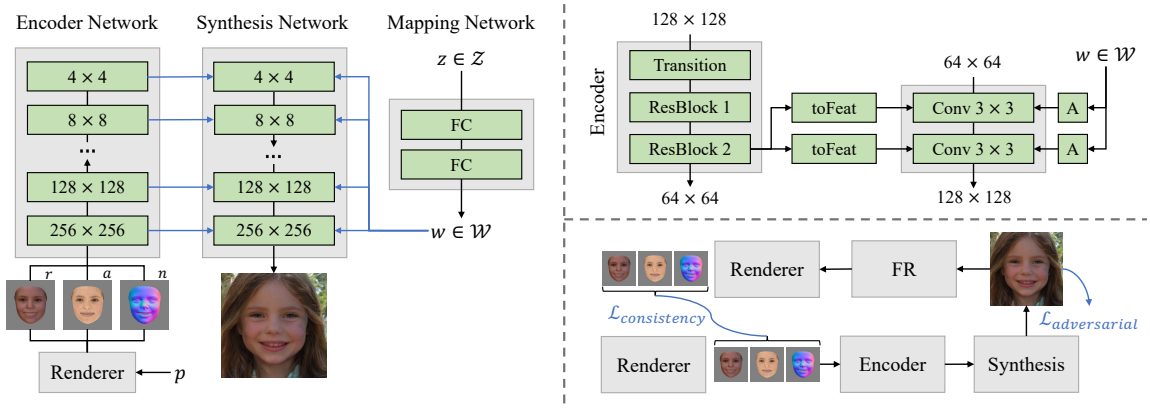


Figure 4.1: We illustrate our overall model architecture (left), a detailed breakdown of feature injection (top right) and our training scheme (bottom right).

ing the diversity of samples produced by G. Furthermore, 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. 4.4 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 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.

4.1.2 *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. 4.3, 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. 4.3, following the intuition that r is always a close enough approximation of the real face for **FR**:

$$\begin{aligned} \mathcal{L}^{*}_{\text{consistency}} &= \mathbb{E}_{p\sim P(p),x\sim G(\text{rep}(p),z)}\left[d\right] \\ d &= \|\text{rep}(\mathbf{FR}(\alpha x + (1-\alpha)r(p))) - \text{rep}(p)\|_2^2 \end{aligned} \tag{4.5}$$

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. We show that our conditioning method approximates Eq. 3.2 without supervision [6, 7], achieving disentanglement *for free* as an inductive bias of the network architecture.

4.2.1 *Conditioning Feature Maps*

We demonstrate our approach upon the common StyleGAN2 architecture [2]. We follow their architecture 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:

$$
\begin{aligned}
e_i &= \begin{cases} E_0(\text{rep}(p)) & i = 0 \\ E_i(e_{i-1}) & i \neq 0 \end{cases} \\
c_{2i} &= \text{toFeat}_{2i}(e_i) \\
c_{2i+1} &= \text{toFeat}_{2i+1}(e_i)
\end{aligned}
\tag{4.6}
$$

We implement E_i as a sequence of a transition layer and two residual blocks (Fig. 4.1). "toFeat" is implemented by a 1×1 convolution [33] with optional downsampling [14] and leaky ReLU activation [34].

4.2.2 *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 [2] 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. For this purpose, 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 [7]. 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 sample quality [6, 7].

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 occasionally notice obvious discrepancies between x and r in previous work [7], and information loss might be a cause.

Our conditioning method avoids both problems. We show in the following section that feature injection 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.

4.2.3 *Disentanglement Analysis*

To simplify analysis, we omit various details from the StyleGAN2 [2] generator (weight demodulation, noise injection, equalized learning rate, etc.). We formulate each layer l_i of the synthesis network as:

$$l_i(p,z) = \mathbf{W}_i * [c_{n-i}(p); s_i(z) \odot \sigma(l_{i-1}(p,z))] + \mathbf{B}_i \tag{4.7}$$

$\mathbf{W}_i$ 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}:

$$\begin{aligned}
\frac{\partial^2 l_i}{\partial_z \partial c_{n-i}} &= \frac{\partial}{\partial_z}\left(\frac{\partial}{\partial c_{n-i}}\left(\mathbf{W}_i * [c_{n-i}; s_i \odot \sigma(l_{i-1})] + \mathbf{B}_i\right)\right) \\
&= \frac{\partial}{\partial_z}\left(\mathbf{W}_i * \frac{\partial}{\partial c_{n-i}}[c_{n-i}; s_i \odot \sigma(l_{i-1})]\right) \\
&= \frac{\partial}{\partial_z}\left(\mathbf{W}_i * [I; 0]\right) \\
&= 0
\end{aligned} \tag{4.8}$$

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 $\sigma(l_{i-1})$:

$$\begin{aligned}
\frac{\partial^2 l_i}{\partial z \partial p} &= \frac{\partial^2 l_i}{\partial z \partial \sigma(l_{i-1})}\frac{\partial \sigma(l_{i-1})}{\partial p} \\
&= \left(\mathbf{W}_i * \left[0; \frac{\partial s_i}{\partial z}\right]\right)\frac{\partial \sigma(l_{i-1})}{\partial p}
\end{aligned} \tag{4.9}$$

We examine the behavior of variation in p:

$$\begin{aligned}
&\frac{\partial \sigma(l_{i-1})}{\partial p} = \frac{\partial \sigma(l_{i-1})}{\partial l_{i-1}}\frac{\partial l_{i-1}}{\partial p} \\
&= \frac{\partial \sigma(l_{i-1})}{\partial l_{i-1}}\left(\mathbf{W}_{i-1} * \left[\frac{\partial c_{n-i+1}}{\partial p}; s_{i-1} \odot \frac{\partial \sigma(l_{i-2})}{\partial p}\right]\right)
\end{aligned} \tag{4.10}$$

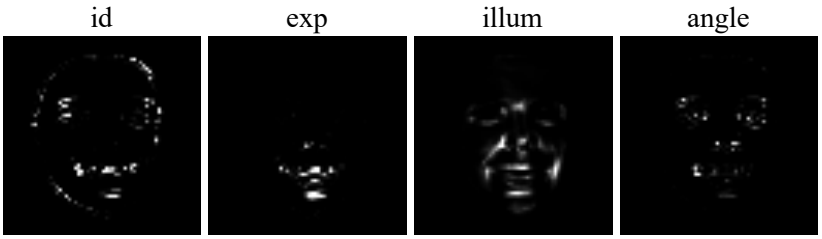


Figure 4.2: Finite difference approximation of the partial derivative of the injected 3DMM render features w.r.t. the 3DMM parameters $\frac{\partial c}{\partial p}$. We can see that the derivative maps are sparse, with the variation in c depicted in small white regions, indicating that disentanglement is mostly successful.

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. 4.2). 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 z's.

4.3 IMPLEMENTATION DETAILS

We implement our model on top of the official StyleGAN2 [2] and the PyTorch release of Deep3DRecon [12]. **FR** and **RDR** are both part of Deep3DRecon [12] and G and D are part of StyleGAN2 [2]. We use the dataset tool provided in Deep3DRecon [12] to realign FFHQ [14] so that image x aligns with 3DMM representation rep.

4.3.1 *StyleGAN2 Backbone*

We follow the latest findings in StyleGAN3 [3] and omit several insignificant details to simplify StyleGAN2 [2]. We remove mixing regularization and path length regularization. The depth of the mapping network is decreased to 2, as recommended by Karras et al. It is also noticed that decreasing the dimensionality of z while maintaining the dimensions of w is beneficial [35]. Therefore, we reduce the dimensions of z to 64. All details are otherwise unchanged, including the network architecture, equalized learning rate, minibatch standard deviation, weight (de)modulation, lazy regularization, bilinear resampling, and exponential moving average of the generator weights.

4.3.2 *Face Reconstruction and Differentiable Renderer*

We use the pretrained checkpoint provided by Deng et al. [12] for **FR**. This updated checkpoint was trained on an augmented dataset that includes FFHQ [14] and shows slight performance improvement over the TensorFlow release of Deep3DRecon. We use the differentiable renderer **RDR** that comes with the checkpoint for **FR** from the same code

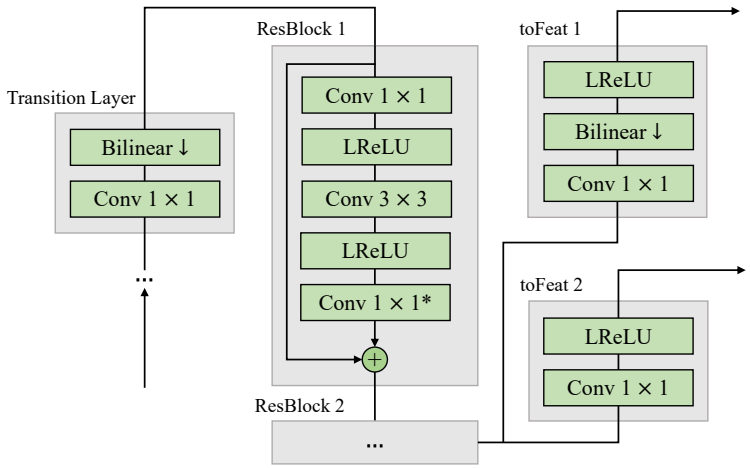


Figure 4.3: The detailed breakdown of a general stage of E.

repository. This renderer uses the Basel Face Model from 2009 [36] as the 3DMM parametric model for face modeling, and nvdiffrast [37] for rasterization. We modify **RDR** so it outputs a and n along with r. The renderer is otherwise unchanged.

4.3.3 *Encoder Architecture*

Figure 4.3 depicts the internal structure of a general stage (every stage other than the highest resolution stage and the 4×4 stage) of our encoder E. Following recent advances in network architecture [38, 39], our ResNet [40] design of E differs from the architecture of D [2] in several ways.

GENERAL STAGE We notice that the two architectural changes in [38] that lead to most performance boost are separate downsampling layers and fewer activations. Thus, we move the skip branch of the transition residual block up to the stem as a transition layer, and remove all activations in the residual block unless they are between two consecutive convolutional layers. We use leaky ReLU activation with $\alpha = 0.2$, and bilinear downsampling instead of strided convolution [2, 14]. We use the 1-3-1 bottleneck residual block as it is more efficient than the 3-3 block [40]. The final convolutional layer (marked by *) in the residual block is initialized to 0 [41], and this eliminates the need for normalization or residual rescaling [2]. We apply equalized learning rate to all convolutional layers.

SPECIALIZATION We remove bilinear downsampling from the transition layer of the highest resolution stage; it is otherwise identical to a general stage. Since the 4×4 stage of the synthesis network contains only one synthesis layer, we place one toFeat layer without leaky ReLU in the 4×4 stage of E accordingly.

4.3.4 *Training Procedure*

Following the StyleGAN family [2, 3, 14], we adopt the non-saturating loss [1] and R1 gradient penalty [42] as the loss function for GAN training. We additional append our $\mathcal{L}_{\text{consistency}}$, resulting in the following objectives:

$$\begin{aligned}\mathcal{L}_D = &-\mathbb{E}_{p,z}\left[\log(1 - D(G(\text{rep}(p), z)))\right] - \\ &\mathbb{E}_x\left[\log(D(x))\right] + \frac{\gamma}{2}\mathbb{E}_x\left[\|\nabla D(x)\|_2^2\right]\end{aligned} \tag{4.11}$$

$$\mathcal{L}_G = -\mathbb{E}_{p,z}\left[\log(D(G(\text{rep}(p), z)))\right] + \lambda\mathcal{L}_{\text{consistency}} \tag{4.12}$$

We closely follow the training configurations of the baseline model in Karras et al. [43] and set $\gamma = 1$. The batch size is set to 64 and the group size of minibatch standard deviation is set to 8. We empirically set $\lambda = 20$ and the length of progressive blending to $k = 2 \times 10^6$. The learning rate of both G and D is set to 2.5×10^{-3}. We train our model until D sees 25M real images [2, 3, 14].

Instead of approximating the distribution $P(p)$ using a VAE [6], we simply use its empirical distribution when sampling $p \sim P(p)$ and find this to be sufficient given our 3DMM representation.

5

EXPERIMENTS

5.1 EXPERIMENTAL SETUP

DATA We use FFHQ [14] at 256×256 resolution to generate our training data. To preprocess the data, we follow the method of Deng et al. [12]: We detect facial landmarks in all FFHQ images with MTCNN [44] 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.

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 [7], along with other frontalization methods [45–49]. 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.

5.2 QUALITATIVE COMPARISON

5.2.1 *Controlled Generation*

Our model achieves highly controllable generation while preserving StyleGAN's ability to generate highly photorealistic images (Fig. 5.1). We can see that our model can produce photorealistic faces with diverse races, genders, and ages. It also shows effective control over each of the 3DMM attributes. Particularly, we use the same three persons as our base images for all attribute edits; this verifies that our model can perform robust generation with high quality.

Our model also demonstrates high consistency when varying either p or z, while holding the other constant. Fig. 5.2 *left* compares the images generated by our model conditioned on the same p but different z's. The identity, expression, pose, and illumination are preserved while all other attributes can be modified. This means there is little overlap between attributes controlled by p and z, and our model gains control over target attributes. On the other hand, when we vary p with fixed z as in Fig. 5.2 *right*. We can see that despite the drastic change in the facial attributes from different p's, the background and clothes re-

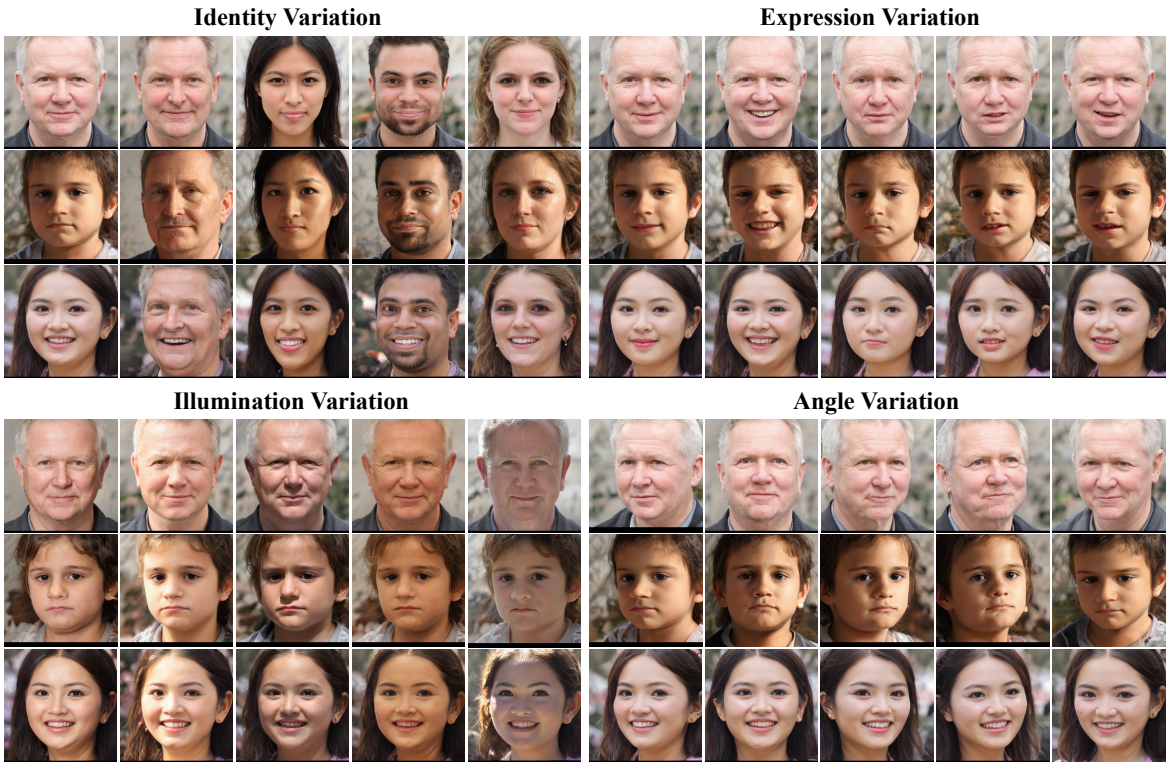


Figure 5.1: Generated face samples with control as output from our model. To attempt to reduce any impression of cherry picking, we use *the same* three input faces for each 3DMM attribute across five edit (columns). While some unwanted variation remains, identity, expression, illumination, and angle are controlled with high fidelity and no apparent visual artifacts.

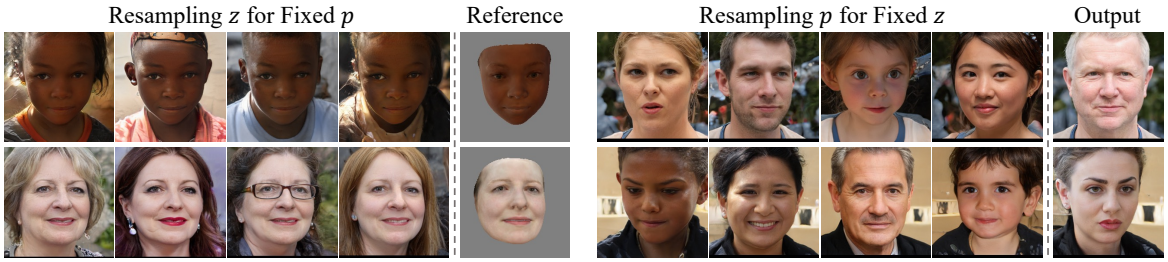


Figure 5.2: *Left.* Resampling the noise vector z with the same set of 3DMM coefficients p shows high facial consistency while other unsupervised factors such as hair, hat, eyeglasses, and background vary. *Right.* Resampling p with the same z shows high consistency in unsupervised factors while the face completely changes.

main largely consistent with the same z. This is another proof that the z vector has a good control of the attributes that are not controlled by p.

5.2.2 *Real Image Inversion and Editing*

Following [7], we test our model's ability to embed real images into its latent space and perform disentangled editing (Fig. 5.3). On zooming, we see that our model produces the sharpest images and that they align closely with the target references from the 3DMM

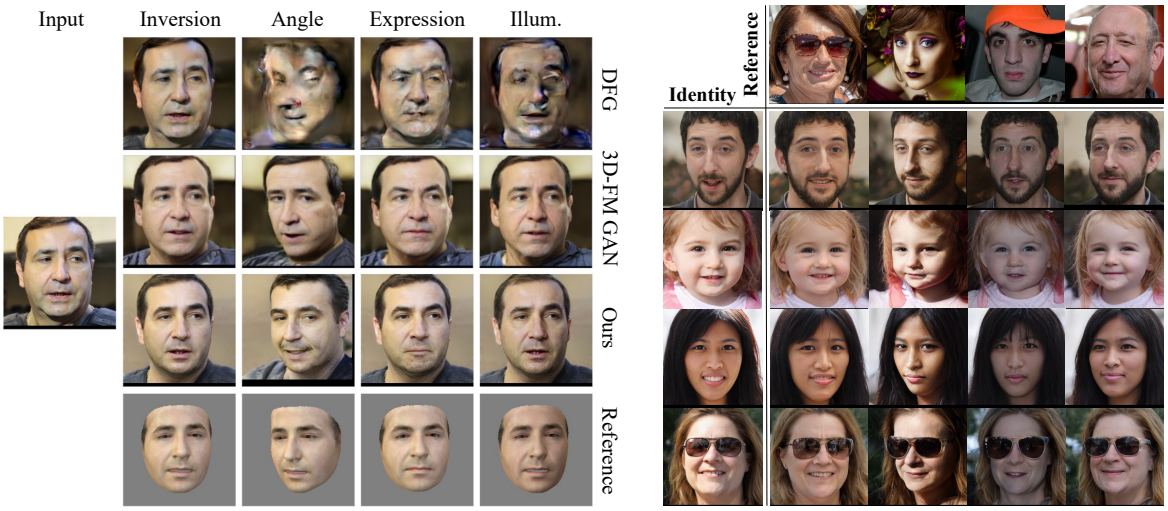


Figure 5.3: *Left.* Competitive real face editing performance demonstrated by our model in both inversion and individual attribute edits, generating faithful iamges to 3DMM renderer outputs and the original image. *Right.* Reference-based generation results. We extract the expression, illumination, and angle coefficients from the reference images (first row) and apply them to randomly generated images (first column).

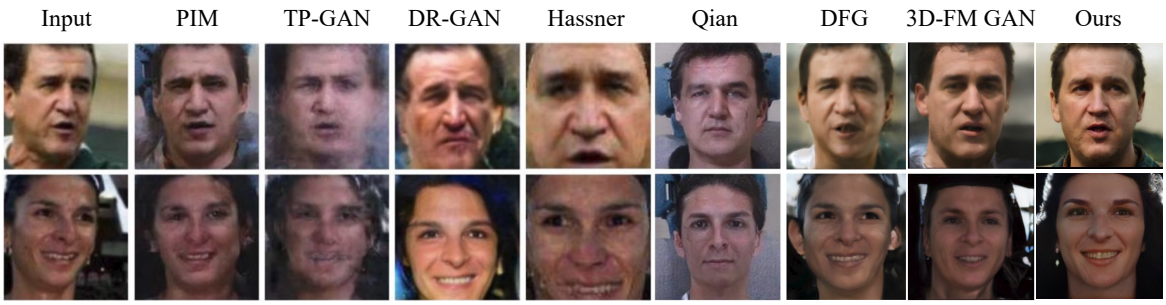


Figure 5.4: Face frontalization comparisons with DiscoFaceGAN (DFG), 3D-FM GAN and other models on LFW images [50]. Our model achieves a good balance between image fidelity and frontal pose positions.

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. 5.4). Our model significantly improves frontalized quality against most methods, and compared with the state-of-the-art face manipulation models [6, 7], our model produces better identity-preserved faces in a more precise frontal view.

5.2.3 *Feature Granularity*

To inspect the impact of feature variability across the layers of the decoder, we inspect the impact of swapping features across images with the same p. In Fig. 5.5, we randomly pick a 3DMM coefficient vector p and randomly sample z's to generate three images (the same images for Source A and Source B). Following StyleGAN [14], we replace some of the style vectors w^+ of images from Source A by the corresponding style vectors of images from

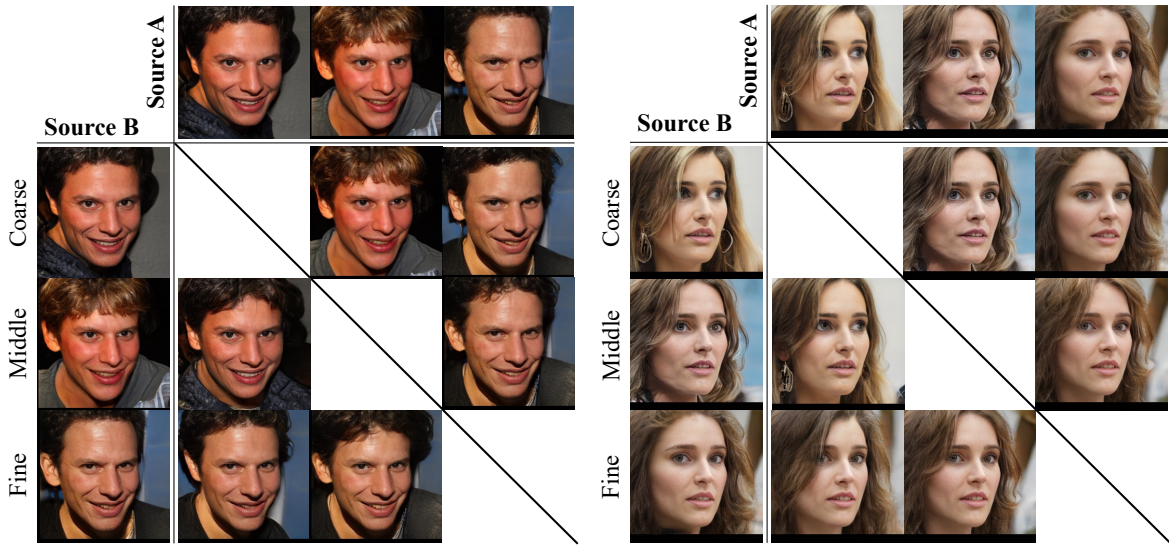


Figure 5.5: Style mixing results at different scales. Using the same three images for Source A and Source B, we replace the style vectors of images from Source A by the style vectors of images from Source B at coarse resolutions (4×4 - 8×8), middle resolutions (16×16 - 32×32), and fine resolutions (64×64 - 256×256).

Source B at coarse, middle, and fine scales. As p is the same, the overall face region will not change significantly.

At coarse scale, there is no visible change to the images from Source A. This is expected as the high-level attributes of the image are supposed to be determined by the p vector. At middle scale, the images from Source A remain mostly unchanged except finer facial features such as the hair now resemble those in the image from Source B. At fine scale, the images from Source A undergo more significant changes where the color scheme that affects the background, clothes, hair color, and skin color now resembles those in the image from Source B. This experiment indicates that each subset of the style vectors w^+ controls a different set of features in the generated image. We also notice that attributes controlled by p remain unchanged at any scale, which means our model's p space and z space are well separated.

5.3 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) [31] and Precision and Recall (P&R) [51, 52] 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. 5.1). Precision and recall indicate that our model has achieved near-StyleGAN-level image generation results while controlling and disentangling facial attributes.

Table 5.1: 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	4.51	0.574	0.485
DiscoFaceGAN	12.9	-	-
3D-FM GAN	12.2	-	-

5.3.1 *Disentanglement Score*

Introduced in DiscoFaceGAN, *disentanglement score* 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.

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 [12]. 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:

$$DS(u_i) = \prod_{j, j \neq i} \frac{\sigma_{u_i}}{\sigma_{u_j}} \tag{5.1}$$

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 5.2).

Table 5.2: Disentanglement Score comparisons with DiscoFaceGAN.

Method	$DS_{\text{id}} \uparrow$	$DS_{\text{exp}} \uparrow$	$DS_{\text{illum}} \uparrow$	$DS_{\text{angle}} \uparrow$
DiscoFaceGAN	0.371	1.64	47.9	829
Ours	**1.02**	**3.22**	**48.7**	**1245**

5.3.2 *Disentanglement, Completeness, and Informativeness*

DCI is a metric introduced in StyleSpace [11]. 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 both models conditioned on these vectors. We then annotate the images by 8 binary classifiers trained on CelebA [14] 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 5.3). 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.

Table 5.3: DCI metric comparisons.

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

5.4 ABLATION STUDY

We modify our model in three different ways to investigate the effect of our proposed methods. We denote our untouched model as **Config-A**. All ablation experiments are conducted on the 128×128 version of FFHQ [14], and all ablation models are trained on 5M real images.

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, 7, 10] do not condition D; it is unclear whether this is an intentional design choice. To our surprise, conditioning D leads to significantly worse FID [31], contradicting the common belief that conditioning is always beneficial [53]. 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. 4.4 proposed by Liu et al. [7]. As expected, this change leads to inferior FID. We notice

Table 5.4: FID and Disentanglement Score comparisons between all ablation experiments.

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	0.412	1.25	23.2	690

that 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. 4.4 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 [2, 53]. 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}.

5.5 INTERACTIVE VISUALIZATION

We develop an interactive program to visualize the controllable face generation and semantic disentanglement of our model. Given an index from the FFHQ dataset, the program retrieves the corresponding 3DMM coefficients p and randomly samples a noise vector z to generate a face image using our trained generator G and a 3DMM reference using **RDR**. Users can then edit the generated face image in real-time, manipulating identity, expression, illumination, and angle through a user-friendly interface. While the 3DMM angle coefficients are highly interpretable and can be directly modified using a 3D angle, the identity, expression, and illumination coefficients require indirect modifications. The program offers two ways to edit these facial attributes: users can either scale the facial attribute coefficients or replace them with those from a reference image.

COEFFICIENT SCALING To modify the identity, expression, and illumination coefficients, users first specify a scaling factor between -1 and 2 using a slider bar. The program

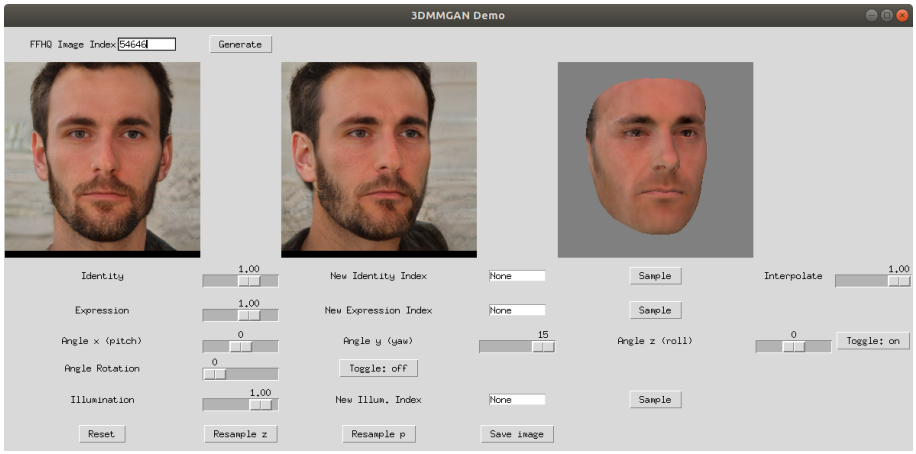


Figure 5.6: Our interactive tool visualizes controllable face generation and editing, demonstrating the semantic disentanglement of our model. Users can manipulate identity, expression, illumination, and angle on the generated faces, and resample either p or z, while keeping the other fixed.

then applies the scaling factor to the corresponding facial attribute and re-generates the face image using the perturbed p and the same z. This approach enables users to visualize a smooth transition of facial attributes. Specifically, increasing the scaling factor magnifies the existing facial attributes, deviating further from the mean FFHQ face, while decreasing the scaling factor shifts the facial attributes closer to the mean face. Negative scaling factors push the facial attributes further towards the opposite direction.

COEFFICIENT REFERENCING To modify the identity, expression, and illumination of generated faces, users can also replace the existing coefficients with those from a reference image. Given another index from the FFHQ dataset, the program retrieves the corresponding 3DMM coefficients for the chosen facial attribute and re-generates the face image using the partially replaced p and the same z. Compared to coefficient scaling, which may represent unrealistic or unnatural facial attributes, referencing real face coefficients results in more realistic changes. Furthermore, the program includes a feature that interpolates between the base and reference facial identity coefficients, highlighting the continuity and smoothness of the 3DMM parameter space.

p AND z RESAMPLING In addition to coefficient editing, the program also allows users to resample either the 3DMM coefficients p or the noise vector z, while holding the other variable fixed. This demonstrates the disentangled control between p and z, as discussed in Section 5.2.1: while p controls identity, expression, illumination, and angle, z controls other attributes such as background and hair that are not part of the 3DMM parametrization.

6

CONCLUSION AND FUTURE WORK

We present a novel conditional model derived from a mathematical framework for 3DMM-conditioned face generation. Our model shows strong performance in both quality and controllability, effectively eliminating the trade-off between these two factors and thereby making control "tax free". Furthermore, our mathematical framework can be applied to future explorations in conditional generation, allowing future investigators to analyze other parametric models in a rigorous manner.

While our proposed model generates photorealistic and controllable deepfake faces, the early work of this project aims to develop a deepfake detection system for social verification settings, where multiple recordings of a public event are analyzed to identify fabricated videos [54]. Prior research has explored frame-by-frame frontalization and 3DMM-based classification for this task, both of which require a well-defined latent or parameter space. In contrast, our model simultaneously incorporates a latent space and a 3DMM parameter space during training, making frontalization and 3DMM coefficient classification natural byproducts. The potential of our model offers new avenues for deepfake detection, particularly upon GAN inversion, where a more sophisticated latent code that provides meanings in both $\mathcal{W}^+$ and the 3DMM parameter space can be obtained.

Despite the strengths of our model, there are certain limitations that should be considered. Specifically, our model is not explicitly designed for image editing, unlike the 3D-FM GAN [7]. As a result, the faces generated by our model exhibit a trade-off between inversion accuracy and editability, same as what has been observed in StyleGAN [2, 3, 14]. To address this issue, future work could explore image editing techniques that achieve a more optimal balance between inversion accuracy and editability for our model. This would further enhance the interpretability of the latent space and hence our model's capability to detect deepfakes.

BIBLIOGRAPHY

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, *Generative adversarial networks*, 2014. arXiv: 1406.2661 [stat.ML].

[2] 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/CVF conference on computer vision and pattern recognition*, 2020, pp. 8110–8119.

[3] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen and T. Aila, 'Alias-free generative adversarial networks,' *Advances in Neural Information Processing Systems*, vol. 34, pp. 852–863, 2021.

[4] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. Denoyer and M. Ranzato, 'Fader networks: Manipulating images by sliding attributes,' *CoRR*, vol. abs/1706.00409, 2017. arXiv: 1706.00409. [Online]. Available: http://arxiv.org/abs/1706.00409.

[5] P. Paysan, R. Knothe, B. Amberg, S. Romdhani and T. 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*, Ieee, 2009, pp. 296–301.

[6] Y. Deng, J. Yang, D. Chen, F. Wen and X. Tong, *Disentangled and controllable face image generation via 3d imitative-contrastive learning*, 2020. DOI: 10.48550/ARXIV.2004.11660. [Online]. Available: https://arxiv.org/abs/2004.11660.

[7] Y. Liu, Z. Shu, Y. Li, Z. Lin, R. Zhang and S. Kung, '3d-fm gan: Towards 3d-controllable face manipulation,' in *European Conference on Computer Vision*, Springer, 2022, pp. 107–125.

[8] A. Tewari, M. Elgharib, G. Bharaj, F. Bernard, H. Seidel, P. Pérez, M. Zollhöfer and C. Theobalt, 'Stylerig: Rigging stylegan for 3d control over portrait images,' *CoRR*, vol. abs/2004.00121, 2020. arXiv: 2004.00121. [Online]. Available: https://arxiv.org/abs/2004.00121.

[9] R. Abdal, P. Zhu, N. J. Mitra and P. Wonka, 'Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows,' *ACM Transactions on Graphics (ToG)*, vol. 40, no. 3, pp. 1–21, 2021.

[10] P. Ghosh, P. S. Gupta, R. Uziel, A. Ranjan, M. J. Black and T. Bolkart, 'Gif: Generative interpretable faces,' in *2020 International Conference on 3D Vision (3DV)*, IEEE, 2020, pp. 868–878.

[11] Z. Wu, D. Lischinski and E. Shechtman, 'Stylespace analysis: Disentangled controls for stylegan image generation,' in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 12 863–12 872.

[12] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia and X. 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*, 2019, pp. 0–0.

[13] T. Karras, T. Aila, S. Laine and J. Lehtinen, 'Progressive growing of gans for improved quality, stability, and variation,' *CoRR*, vol. abs/1710.10196, 2017. arXiv: 1710.10196. [Online]. Available: http://arxiv.org/abs/1710.10196.

[14] T. Karras, S. Laine and T. Aila, 'A style-based generator architecture for generative adversarial networks,' in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[15] Y. Shen, C. Yang, X. Tang and B. Zhou, 'Interfacegan: Interpreting the disentangled face representation learned by gans,' *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 4, pp. 2004–2018, 2020.

[16] Y. Shen and B. Zhou, 'Closed-form factorization of latent semantics in gans,' in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 1532–1540.

[17] E.-T. Le, E. Bartrum and I. Kokkinos, 'Stylemorph: Disentangled 3d-aware image synthesis with a 3d morphable styleGAN,' in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=Ojpb1y8jflw.

[18] Y. Guo, J. Cai, B. Jiang, J. 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*, vol. 41, no. 6, pp. 1294–1307, 2018.

[19] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi and R. Ng, 'Nerf: Representing scenes as neural radiance fields for view synthesis,' *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[20] J. Gu, L. Liu, P. Wang and C. Theobalt, 'Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis,' *arXiv preprint arXiv:2110.08985*, 2021.

[21] E. R. Chan, M. Monteiro, P. Kellnhofer, J. Wu and G. 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*, 2021, pp. 5799–5809.

[22] K. Sun, S. Wu, Z. Huang, N. Zhang, Q. Wang and H. Li, 'Controllable 3d face synthesis with conditional generative occupancy fields,' *arXiv preprint arXiv:2206.08361*, 2022.

[23] Y. Bengio, A. Courville and P. Vincent, 'Representation learning: A review and new perspectives,' *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[24] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever and P. Abbeel, 'Infogan: Interpretable representation learning by information maximizing generative adversarial nets,' *CoRR*, vol. abs/1606.03657, 2016. arXiv: 1606.03657. [Online]. Available: http://arxiv.org/abs/1606.03657.

[25] Z. Lin, K. Thekumparampil, G. Fanti and S. Oh, 'Infogan-cr and modelcentrality: Self-supervised model training and selection for disentangling gans,' in *international conference on machine learning*, PMLR, 2020, pp. 6127–6139.

[26] J. Pamuła, *Progressive training of gans with mutual information penalty.* [Online]. Available: https://github.com/jonasz/progressive_infogan.

[27] W. Peebles, J. Peebles, J.-Y. Zhu, A. A. Efros and A. Torralba, 'The hessian penalty: A weak prior for unsupervised disentanglement,' in *Proceedings of European Conference on Computer Vision (ECCV)*, 2020.

[28] A. Tewari, M. Elgharib, M. B. R., F. Bernard, H.-P. Seidel, P. Pérez, M. Zollhöfer and C. Theobalt, *Pie: Portrait image embedding for semantic control*, 2020. DOI: 10.48550/ARXIV.2009.09485. [Online]. Available: https://arxiv.org/abs/2009.09485.

[29] W. Nie, T. Karras, A. Garg, S. Debnath, A. Patney, A. B. Patel and A. Anandkumar, 'Semi-supervised stylegan for disentanglement learning,' in *Proceedings of the 37th International Conference on Machine Learning*, 2020, pp. 7360–7369.

[30] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf and O. Bachem, 'Challenging common assumptions in the unsupervised learning of disentangled representations,' in *international conference on machine learning*, PMLR, 2019, pp. 4114–4124.

[31] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler and S. Hochreiter, 'Gans trained by a two time-scale update rule converge to a local nash equilibrium,' *Advances in neural information processing systems*, vol. 30, 2017.

[32] I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[33] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, 'Gradient-based learning applied to document recognition,' *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. DOI: 10.1109/5.726791.

[34] A. L. Maas, A. Y. Hannun, A. Y. Ng *et al.*, 'Rectifier nonlinearities improve neural network acoustic models.'

[35] A. Sauer, K. Schwarz and A. Geiger, 'Stylegan-xl: Scaling stylegan to large diverse datasets,' in *ACM SIGGRAPH 2022 conference proceedings*, 2022, pp. 1–10.

[36] *A 3D Face Model for Pose and Illumination Invariant Face Recognition*, IEEE, Genova, Italy, 2009.

[37] S. Laine, J. Hellsten, T. Karras, Y. Seol, J. Lehtinen and T. Aila, 'Modular primitives for high-performance differentiable rendering,' *ACM Transactions on Graphics*, vol. 39, no. 6, 2020.

[38] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell and S. Xie, 'A convnet for the 2020s,' in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 976–11 986.

[39] W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng and S. Yan, 'Metaformer is actually what you need for vision,' in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 819–10 829.

[40] K. He, X. Zhang, S. Ren and J. Sun, 'Deep residual learning for image recognition,' in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[41] H. Zhang, Y. N. Dauphin and T. Ma, 'Fixup initialization: Residual learning without normalization,' *arXiv preprint arXiv:1901.09321*, 2019.

[42] L. Mescheder, A. Geiger and S. Nowozin, 'Which training methods for gans do actually converge?' In *International conference on machine learning*, PMLR, 2018, pp. 3481–3490.

[43] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen and T. Aila, 'Training generative adversarial networks with limited data,' *Advances in neural information processing systems*, vol. 33, pp. 12 104–12 114, 2020.

[44] K. Zhang, Z. Zhang, Z. Li and Y. Qiao, 'Joint face detection and alignment using multi-task cascaded convolutional networks,' *CoRR*, vol. abs/1604.02878, 2016. arXiv: 1604.02878. [Online]. Available: http://arxiv.org/abs/1604.02878.

[45] J. Zhao, Y. Cheng, Y. Xu, L. Xiong, J. Li, F. Zhao, K. Jayashree, S. Pranata, S. Shen, J. Xing *et al.*, 'Towards pose invariant face recognition in the wild,' in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2207–2216.

[46] R. Huang, S. Zhang, T. Li and R. 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*, 2017, pp. 2439–2448.

[47] L. Tran, X. Yin and X. Liu, 'Disentangled representation learning gan for pose-invariant face recognition,' in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1415–1424.

[48] T. Hassner, S. Harel, E. Paz and R. Enbar, 'Effective face frontalization in unconstrained images,' in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4295–4304.

[49] Y. Qian, W. Deng and J. 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*, 2019, pp. 9851–9858.

[50] G. B. Huang, M. Mattar, T. Berg and E. 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.

[51] T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen and T. Aila, 'Improved precision and recall metric for assessing generative models,' *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[52] M. S. Sajjadi, O. Bachem, M. Lucic, O. Bousquet and S. Gelly, 'Assessing generative models via precision and recall,' *Advances in neural information processing systems*, vol. 31, 2018.

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

[54] E. Tursman, M. George, S. Kamara and J. Tompkin, 'Towards untrusted social video verification to combat deepfakes via face geometry consistency,' in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 2784–2793. DOI: 10.1109/CVPRW50498.2020.00335.