Disentangled and Controllable Face Image Generation via 3D Imitative-Contrastive Learning

Yu Deng*[1,2] Jiaolong Yang[2] Dong Chen[2] Fang Wen[2] Xin Tong[2]
[1]Tsinghua University [2]Microsoft Research Asia
{t-yudeng,jiaoyan,doch,fangwen,xtong}@microsoft.com

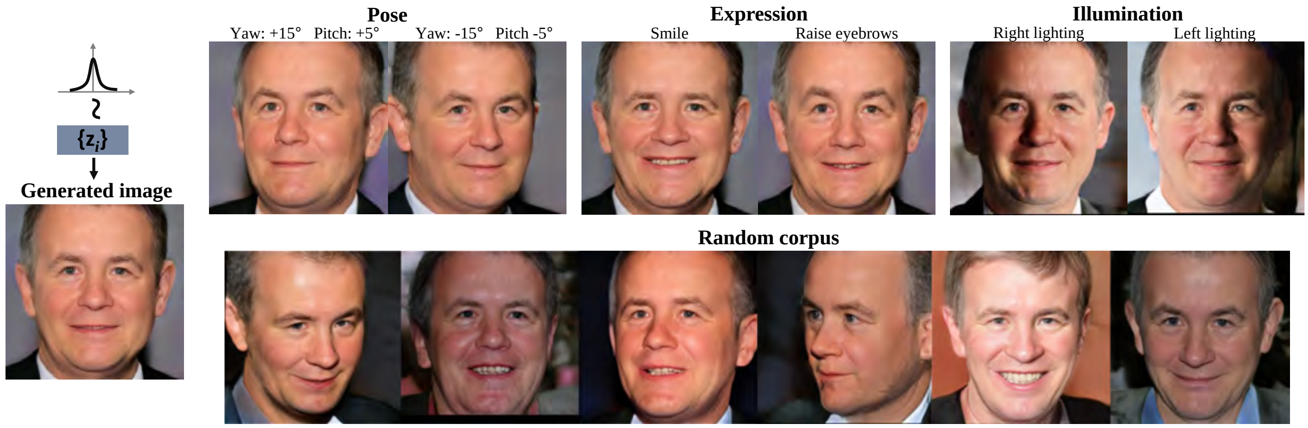

Figure 1: This paper presents ***DiscoFaceGAN*** that generates realistic face images of virtual people with independent latent variables of identity, expression, pose, and illumination. The latent space is interpretable and highly disentangled, which allows precise control of the targeted images (*e.g.*, degree of each pose angle, lighting intensity and direction), as shown in the top row. The bottom row shows the generated images when we keep the identity and randomize other properties. The faces generated by our method are not any real person in the world.

Abstract

*We propose **DiscoFaceGAN**, an approach for face image generation with **DIS**entangled, precisely-**CO**ntrollable latent representations for identity of non-existing people, expression, pose, and illumination. We embed 3D priors into adversarial learning and train the network to imitate the image formation of an analytic 3D face deformation and rendering process. To deal with the generation freedom induced by the domain gap between real and rendered faces, we further introduce contrastive learning to promote disentanglement by comparing pairs of generated images. Experiments show that through our imitative-contrastive learning, the factor variations are very well disentangled and the properties of a generated face can be precisely controlled. We also analyze the learned latent space and present several meaningful properties supporting factor disentanglement. Our method can also be used to embed real images into the disentangled latent space. We hope our method could provide new understandings of the relationship between physical properties and deep image synthesis.* [1]

*This work was done when Yu Deng was an intern at MSRA.
[1]Code available here.

1. Introduction

Face image synthesis has achieved tremendous success in the past few years with the rapid advance of Generative Adversarial Networks (GANs) [14]. State-of-the art GAN models, such as the recent StyleGAN [23], can generate high-fidelity virtual face images that are sometimes even hard to distinguish from real ones.

Compared to the vast body of works devoted to improving the image generation quality and tailoring GANs for various applications, synthesizing face images *de novo* with multiple disentangled latent spaces characterizing different properties of a face image is still not well investigated. Such a disentangled latent representation is desirable for constrained face image generation (*e.g.*, random identities with specific illuminations or poses). It can also derive a disentangled representation of a real image by embedding it into the learned feature space. A seminal GAN research for disentangled image generation is InfoGAN [6], where the representation disentanglement is learned in an unsupervised manner via maximizing the mutual information between the latent variables and the observation. However, it has been shown that without any prior or weak supervision, there is no guarantee that each latent variable contains a semantically-meaningful factor of variation [30, 7].

In this paper, we investigate synthesizing face images of virtual people with independent latent variables for identity, expression, pose, lighting, and an additional noise. To gain predictable controllability on the former four variables, we translate them to the coefficients of parametric models through training a set of Variational Autoencourders (VAE). We incorporate priors from 3D Morphable Face Models (3DMM) [4, 33] and an analytic rendering procedure into adversarial learning. A set of *imitative losses* is introduced which enforces the generator to imitate the explainable image rendering process, thus generating face properties characterized by the latent variables. However, the domain gap between real and rendered faces gives rise to a certain generation freedom that is uncontrollable, leading to unsatisfactory disentanglement of factor variations.

To deal with such generation freedom and enhance disentanglement, we further propose a collection of *contrastive losses* for training. We compare pairs of generated images and penalize the appearance difference that is only induced by a set of identical latent variables shared between each pair. This way, the generator is forced to express an independent influence of each latent variable to the final output. We show that these contrastive losses are crucial to achieve complete latent variable disentanglement.

The model we use in this paper is based on the StyleGAN structure [23], though our method can be extended to other GAN models as well. We modify the latent code layer of StyleGAN and equip it with our new loss functions for training. We show that the latent variables can be highly disentangled and the generation can be accurately controlled. Similar to StyleGAN, the faces generated by our method do not correspond to any real person in the world. We further analyze the learned StyleGAN latent space and find some meaningful properties supporting factor disentanglement. Our method can be used to embed real images into the disentangled latent space and we demonstrate this with various experiments.

The contributions of this paper can be summarized as follows. We propose a novel disentangled representation learning scheme for *de novo* face image generation via a imitative-contrastive paradigm leveraging 3D priors. Our method enables precise control of the targeted face properties such as pose, expression, and illumination, achieving flexible and high-quality face image generation that, to our knowledge, cannot be achieved by any previous method. Moreover, we offer several analyses to understand the properties of the disentangled StyleGAN latent space. At last, we demonstrate that our method can be used to project real images into the disentangled latent space for analysis and decomposition.

2. Related Work

We briefly review the literature on disentangled representation learning and face image synthesis as follows.

Disentangled representation learning. Disentangled representation learning (DRL) for face images has been vividly studied in the past. Historical attempts are based on simple bilinear models [46], restricted Boltzmann machines [10, 39], among others. A seminal GAN research along this direction is InfoGAN [6]. However, InfoGAN is known to suffer from training instability [48], and there is no guarantee that each latent variable is semantically meaningful [30, 7]. InfoGAN-CR [29] introduces an additional discriminator to identify the latent code under traversal. SD-GAN [11] applies a discriminator on image pairs to disentangle identity and appearance factors. Very recently, HoloGAN [32] disentangles 3D pose and identity with unsupervised learning using 3D convolutions and rigid feature transformations. DRL with VAEs also received much attention in recent years [26, 48, 18, 5, 25].

Conditional GAN for face synthesis. CGAN [31] has been widely used in face image synthesis tasks especially identity-preserving generation [47, 2, 52, 3, 42]. In a typical CGAN framework, the input to a generator consists of random noises together with some preset conditional factors (*e.g.*, categorical labels or features) as constraints, and an auxiliary classifier/feature extractor is applied to restore the conditional factors from generator outputs. It does not offer a generative modeling of the conditional factors. Later we show that our method can be applied to various face generation tasks handled previously with CGAN frameworks.

Face image embedding and editing with GANs. GANs have seen heavy use in face image manipulation [34, 19, 49, 44, 36, 45, 54]. These methods typically share an encoder-decoder/generator-discriminator paradigm where the encoder embeds images into disentangled latent representations characterizing different facial properties. Our method can also be applied to embed face images into our disentangled latent space, as we will show in the experiments.

3D prior for GANs. Many methods have been proposed to incorporate 3D prior into GAN for face synthesis [52, 43, 24, 8, 12, 35, 13, 32, 50]. Most of them leverages 3DMMs. For example, [24] utilizes 3DMM coefficients extracted from input images as low-frequency feature for frontal face synthesis. [12] and [35] translate rendered 3DMM faces and real face images in a cycle fashion. [24] generates video frames from 3DMM faces for face re-animation. [50] uses 3DMM for portrait reconstruction and pose manipulation. Different from these methods, we only employ 3DMM as priors in the training stage for our imitative-contrastive learning. After training, we do not require a 3DMM model or any rendering procedure.

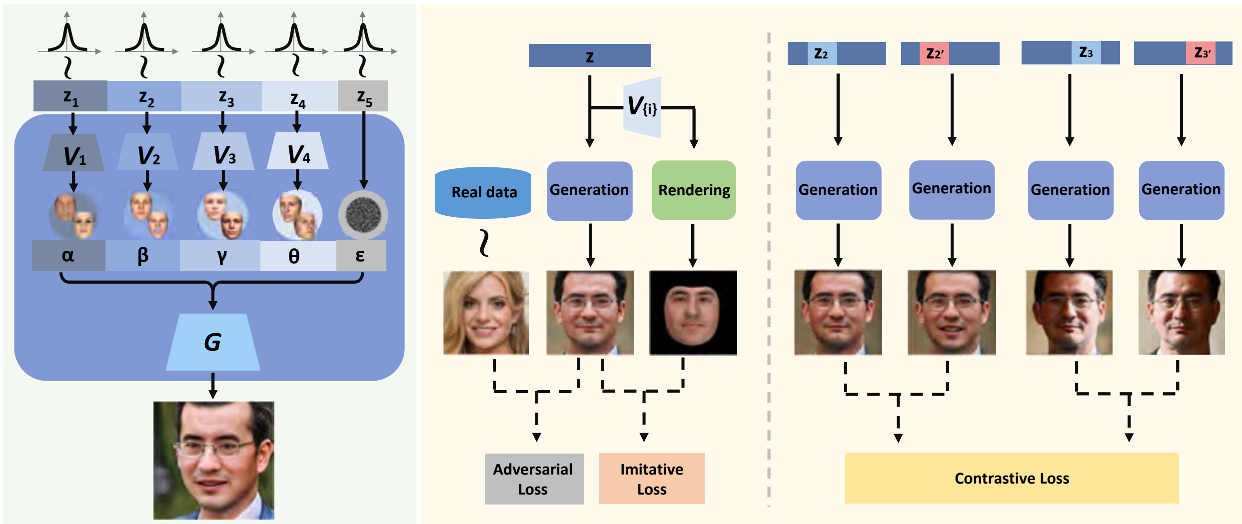

Figure 2: Overview of the ***DiscoFaceGAN***. The left diagram (in green) shows the generation pipeline, and the rest illustrates our training scheme which features three type of losses: adversarial loss, imitative loss, and contrastive loss.

3. Approach

Given a collection of real face images $\mathcal{Y}$, our goal is to train a network G that generates realistic face images x from random noise z, which consists of multiple independent variables $z_i \in \mathbb{R}^{N_i}$, each following the normal distribution. We consider latent variables for five independent factors: identity, expression, illumination, pose, and a random noise accounting for other properties such as background. As in standard GAN, a discriminator D is applied to compete with G. To obtain disentangled and interpretable latent space, we incorporate 3D priors in an imitative-contrastive learning scheme (Fig. 2), described as follows.

3.1. Imitative Learning

To learn how a face image should be generated following the desired properties, we incorporate a 3DMM model [33] and train the generator to imitate the rendered 3D faces. With a 3DMM, the 3D shape $\mathbf{S}$ and texture $\mathbf{T}$ of a face is parameterized as

$$\begin{aligned}\mathbf{S} &= \bar{\mathbf{S}} + \mathbf{B}_{id}\alpha_s + \mathbf{B}_{exp}\beta \\ \mathbf{T} &= \bar{\mathbf{T}} + \mathbf{B}_t\alpha_t\end{aligned} \tag{1}$$

where $\bar{\mathbf{S}}$ and $\bar{\mathbf{T}}$ are the average face shape and texture, $\mathbf{B}_{id}$, $\mathbf{B}_{exp}$, and $\mathbf{B}_t$ are the PCA bases of identity, expression, and texture, respectively, and α_s, β, and α_t are the corresponding 3DMM coefficient vectors. We denote $\alpha \doteq [\alpha_s, \alpha_t]$ as the identity-bearing coefficients. We approximate scene illumination with Spherical Harmonics (SH) [38] parameterized by coefficient vector γ. Face pose is defined as three rotation angles[2] expressed as vector θ. With $\lambda \doteq [\alpha, \beta, \gamma, \theta]$, we can easily obtain a rendered face $\hat{x}$ through a well-established analytic image formation [4].

To enable imitation, we first bridge the z-space to the λ-space. We achieve this by training VAE models on the λ samples extracted from real image set $\mathcal{Y}$. More specifically, we use the 3D face reconstruction network from [9] to obtain the coefficients of all training images and train four simple VAEs for α, β, γ and θ, respectively. After training, we discard the VAE encoders and keep the decoders, denoted as V_i, $i = 1, 2, 3, 4$, for z-space to λ-space mapping.

In our GAN training, we sample $\mathrm{z} = [z_1, \ldots, z_5]$ from standard normal distribution, map it to λ, and feed λ to both the generator G and the renderer to obtain a generated face x and a rendered face $\hat{x}$, respectively. Note that we can input either z or λ into G – in practice we observe no difference between these two options in terms of either visual quality or disentangling efficacy. The benefit of using λ is the ease of face property control since λ is interpretable.

We define the following loss functions on x for imitative learning. First, we enforce x to mimic the identity of $\hat{x}$ perceptually by

$$l_I^{id}(x) = \max(1- < f_{id}(x), f_{id}(\hat{x}) > -\tau, 0), \tag{2}$$

where $f_{id}(\cdot)$ is the deep identity feature from a face recognition network, $< \cdot, \cdot >$ denotes cosine similarity, and τ is a constant margin which we empirically set as 0.3. Since there is an obvious domain gap between rendered 3DMM faces and real ones, we allow a small difference between the features. The face recognition network from [51] is used in this paper for deep identity feature extraction. For expression and pose, we penalize facial landmark differences via

$$l_I^{lm}(x) = \|p(x) - \hat{p}\|^2, \tag{3}$$

[2] We align the images to cancel translation.

where $p(\cdot)$ denotes the landmark positions detected by the 3D face reconstruction network, and $\hat{p}$ is the landmarks of the rendered face obtained trivially. For illumination, we simply minimize the SH coefficient discrepancy by

$$l_I^{sh}(x) = |\gamma(x) - \hat{\gamma}|_1, \tag{4}$$

where $\gamma(\cdot)$ represents the coefficient given by the 3D face reconstruction network, and $\hat{\gamma}$ is the coefficient of $\hat{x}$. Finally, we add a simple loss which enforces the output to mimic the skin color of the rendered face via

$$l_I^{cl}(x) = |c(x) - c(\hat{x})|_1, \tag{5}$$

where $c(\cdot)$ denotes the average color of face region defined by the mask in 3DMM. By using these imitative losses, the generator will learn to generate face images following the identity, expression, pose, and illumination characterized by the corresponding latent variables.

The domain gap issue. Obviously, there is an inevitable domain gap between the rendered 3DMM faces and generated ones. Understanding the effect of this domain gap and judiciously dealing with it is important. On one hand, retaining a legitimate domain gap that is reasonably large is necessary as it avoids the conflict with the adversarial loss and ensures the realism of generated images. It also prevents the generative modeling from being trapped into the small identity subspace of the 3DMM model[3]. On the other hand, however, it may lead to poor factor variation disentanglement (for example, changing expression may lead to unwanted variations of identity and image background, and changing illumination may disturb expression and hair structure; see Fig. 3 and 6).

To understand why this happens, we first symbolize the difference between a generated face x and its rendered counterpart $\hat{x}$ as Δx, *i.e.*, $x = \hat{x} + \Delta x$. In the imitative learning, x is free to deviate from $\hat{x}$ in terms of certain identity characteristics and other image contents beyond face region (*e.g.*, background, hair, and eyewear). As a consequence, Δx has a certain degree of freedom that is *uncontrollable*. We resolve this issue via contrastive learning, to be introduced next.

3.2. Contrastive Learning

To fortify disentanglement, we enforce the invariance of the latent representations for image generation in a contrastive manner: we vary one latent variable while keeping others unchanged, and enforce that the difference on the generated face images relates only to that latent variable. Concretely, we sample pairs of latent code z, z' which differ only at z_k and share the same $z_i, \forall i \neq k$. We compare the generated face images x, x', and then penalize the difference induced by any of z_i but z_k.

[3]The 3DMM we use in this paper is from [33] which is constructed by scans of 200 people.

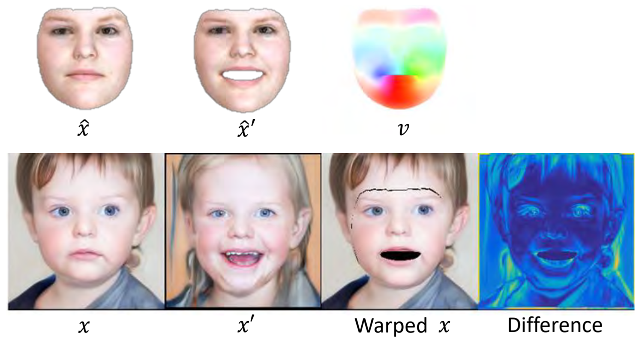

Figure 3: Illustration of the image warping process in our contrastive learning.

To enable such a comparison, we need to find a function $\phi_k(G(\mathrm{z}))$ which is, to the extent possible, invariant to z_k but sensitive to variations of z_i's. In this work, we implement two simple functions for face images. The first one is designed for expression-invariant comparison. Our idea is to restore a neutral expression for x and x' to enable the comparison. However, high-fidelity expression removal *per se* is a challenging problem still being actively studied in GAN-based face image manipulation [37, 13]. To circumvent this issue, we resort to the rendered 3DMM face $\hat{x}$ to get a surrogate flow field for image warping. Such a flow field can be trivially obtained by revising the expression coefficient and rendering another 3DMM face with a neutral expression. In practice, it is unnecessary to warp both x and x'. We simply generate the flow field v from $\hat{x}$ to $\hat{x}'$ and warp x to x' accordingly (see Fig. 3 for an example). We then minimize the image color difference via

$$l_C^{ex}(x, x') = |x(v) - x'|_1, \tag{6}$$

where $x(v)$ is the warped image.

Second, we design two illumination-invariant losses for contrastive learning. Since the pixel color across the whole image can be affected by illumination change, we simply enforce the semantical structure to remain static. We achieve this by minimizing the difference between the face structures of x and x':

$$l_C^{il_1}(x, x') = \|m(x) - m(x')\|^2 + \omega\|p(x) - p(x')\|^2, \tag{7}$$

where $m(\cdot)$ is the hair segmentation probability map obtained from a face parsing network [28], $p(\cdot)$ denotes landmark positions same as in Eq. 3, and ω is a balancing weight. We also apply a deep identity feature loss via

$$l_C^{il_2}(x, x') = 1 - < f_{id}(x), f_{id}(x') > . \tag{8}$$

In this paper, using the above contrastive learning losses regarding expression and illumination can lead to satisfactory disentanglement (we found that pose variations can be well disentangled without need for another contrastive loss).

Figure 4: Face images generated by our *DiscoFaceGAN*. As shown in the figures, the variations of identity, expression, pose and illumination are highly disentangled, and we can precisely control expression, illumination and pose.

Effect of contrastive learning. Following the discussion in Section 3.1, for two rendered faces $\hat{x}$ and $\hat{x}'$ which only (and perfectly) differ at one factor such as expression, both Δx and $\Delta x'$ have certain free variations that are uncontrollable. Therefore, achieving complete disentanglement with imitative learning is difficult, if not impossible. The contrastive learning is an essential complement to imitative learning: it imposes proper constrains on Δx and $\Delta x'$ by explicitly learning the desired differences between x and x', thus leading to enhanced disentanglement.

We empirically find that the contrastive learning also leads to better imitation and more accurate face property control. This is because the pairwise comparison can also suppress imitation noise: any misalignment of pose or expression between x and $\hat{x}$ or between x' and $\hat{x}'$ will incur larger contrastive losses.

4. Experiments

Implementation details. In this paper, we adopt the StyleGAN structure [23] and the FFHQ dataset [23] for training. We train the λ-space VAEs following the schedule of [7], where encoders and decoders of the VAEs are all MLPs with three hidden layers. For StyleGAN, we follow the standard training procedure of the original method except that we 1) remove the normalization operation for input latent variable layer, 2) discard the style-mixing strategy, and 3) train up to image resolution of 256×256 due to time constraint. We first train the network with the adversarial loss as in [23] and our imitative losses until seeing $15M$ real images to obtain reasonable imitation. Then we add contrastive losses into the training process and train the network up to seeing $20M$ real images in total. More training details can be found in the *suppl. material*.

Figure 5: Reference-based generation results where we extract expression, lighting, and pose properties of real images and combine them with randomly generated identities.

4.1. Generation Results

Figure 4 presents some image samples generated by our *DiscoFaceGAN* after training. It can be seen that our method is able to randomly generate high-fidelity face images with a large variant of identities with diverse pose, illumination, and facial expression. More importantly, the variations of identity, expression, pose, and illumination are highly disentangled – when we vary one factor, all others can be well preserved. Furthermore, we can precisely control expression, illumination and pose using the parametric model coefficients for each of them. One more example for precisely controlled generation is given in Fig. 1.

Figure 5 shows that we can generate images of new identities by mimicking the properties of a real reference image. We achieve this by extracting the expression, lighting, and pose parameters from the reference image and combine them with random identity variables for generation.

4.2. Ablation Study

In this section, we train the *DiscoFaceGAN* with different losses to validate the effectiveness of our imitative-contrastive learning scheme. Some typical results are presented in Fig. 6. Obviously, the network cannot generate reasonable face images if we remove the imitation losses. This is because the contrastive losses rely on reasonable imitation, without which they are less meaningful and the network behavior will be unpredictable. On the other hand, without contrastive losses, variations of different factors cannot be fully disentangled. For example, expression and lighting changes may influence certain identity-related characteristics and some other properties such as hair structure. The contrastive losses can also improve the desired preciseness of imitation (*e.g.*, see the mouth-closing status in the last row), leading to more accurate generation control.

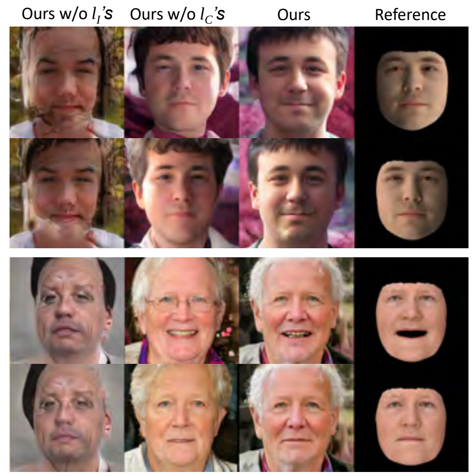

Figure 6: Ablation study of the training losses. The top and bottom two rows show the results when we vary the latent variable for lighting and expression, respectively.

Table 1: Comparison of disentanglement score as well as generation quality.

	Disentanglement ↑				Quality ↓	
	DS_α	DS_β	DS_γ	DS_θ	FID	PPL
3DMM	-				271	-
l_{adv}	0.83	1.98	0.87	0.07	**5.49**	106
$+l'_I s$	**13.4**	37.0	40.4	31.6	9.15	**102**
$+l'_C s$	7.85	**80.4**	**489**	**36.7**	12.9	123

4.3. Quantitative Evaluation

In this section, we evaluate the performance of our *DiscoFaceGAN* quantitatively in terms of disentanglement efficacy as well as generation quality. For the former, several metrics have been proposed in VAE-based disentangled representation learning, such as factor score [25] and mutual information gap [5]. However, these metrics are not suitable for our case. Here we design a simple metric named disentanglement score (DS), described as follows.

Our goal is to measure that when we only vary the latent variable for one single factor, if other factors on the generated images are stable. We denote the four λ-space variables $\alpha, \beta, \gamma, \theta$ as u_i, and we use $u_{\{j\}}$ as the shorthand notation for the variable set $\{u_j | j = 1, \ldots, 4, j \neq i\}$. To measure the disentanglement score for u_i, we first randomly generate 1K sets of $u_{\{j\}}$, and for each $u_{\{j\}}$ we randomly generate 10 u_i. Therefore, we can generate 10K images using the trained network with combinations of u_i and $u_{\{j\}}$. For these images, we re-estimate u_i and $u_{\{j\}}$ using the 3D reconstruction network [9] (for identity we use a face recognition network [51] to extract deep identity feature instead). We calculate the variance of the estimated values for each of the 1K groups, and then average them to obtain σ_{u_i} and σ_{u_j}. We further normalize σ_{u_i} and σ_{u_j} by dividing the variance of the corresponding variable computed on FFHQ. Finally, we measure the disentanglement score via

$$DS_{u_i} = \Pi_{j,j\neq i}\frac{\sigma_{u_i}}{\sigma_{u_j}}, \tag{9}$$

A high DS indicates that when varying a certain factor, only the corresponding property in the generated images is changing ($\sigma_{u_i} > 0$) while other factors remain unchanged ($\sigma_{u_j} \to 0$). Table 1 shows that the imitative learning leads to high factor disentanglement and the contrastive learning further enhances it for expression, illumination, and pose. The disentanglement score for identity decreases with contrastive learning. We found the 3D reconstruction results from the network are slightly unstable when identity changes, which increased the variances of other factors.

To evaluate the quality of image generation, we follow [23] to compute the Fréchet Inception Distances (FID) [17] and the Perceptual Path Lengths (PPL) [23] using 50K and 100K randomly generated images, respectively. Table 1 shows that the FID increases with our method. This is expected as the additional losses added to the adversarial training will inevitably affect the generative modeling. However, we found that the PPL is comparable to the results trained with only the adversarial loss.

5. Latent Space Analysis and Embedding

In this section, we analyze the latent space of our *DiscoFaceGAN*. We show some meaningful properties supporting factor variation disentanglement, based on which we further present a method for embedding and manipulating real face images in the disentangled latent space.

5.1. Analysis of Latent Space

One key ingredients of StyleGAN is the mapping from z-space to $\mathcal{W}$-space, the latter of which relates linearly to the AdaIN [22] parameters that control "styles" (we refer the readers to [23] for more details). Previous studies [41, 1] have shown that certain *direction of changes* in $\mathcal{W}$-space leads to variations of corresponding attributes in generated images. In our case, $\mathcal{W}$ space is mapped from λ space which naturally relates to image attributes. Therefore, we analyze the *direction of changes* in the learned $\mathcal{W}$-space by varying λ variables, and some interesting properties have been found. We will introduce these properties and then provide strong empirical evidences supporting them.

Recall that the input to generator is λ-space variables $\alpha, \beta, \gamma, \theta$ and an additional noise ε. Here we denote these five variables as u_i with $u_5 = \varepsilon$. We use $u_{\{j\}}$ as the shorthand notation for the variable set $\{u_j | j = 1, \ldots, 5, j \neq i\}$, and $w(u_i, u_{\{j\}})$ denotes the $\mathcal{W}$ space variable mapped from u_i and $u_{\{j\}}$. We further denote a unit vector

$$\widehat{\Delta w}(i,a,b) = \frac{w(u_i{=}a, u_{\{j\}}) - w(u_i{=}b, u_{\{j\}})}{\|w(u_i{=}a, u_{\{j\}}) - w(u_i{=}b, u_{\{j\}})\|} \tag{10}$$

Table 2: Cosine similarities of direction of change in $\mathcal{W}$ space. **Top:** changing a factor from a fixed start to a fixed end. **Bottom:** changing a factor with a fixed offset.

	identity	expression	lighting	pose
l_{adv}	0.65± 0.10	0.21± 0.11	0.16± 0.12	0.17± 0.11
Ours	**0.96± 0.02**	**0.82± 0.04**	**0.85± 0.03**	**0.87± 0.03**

	identity	expression	lighting	pose
l_{adv}	0.42± 0.14	0.21± 0.12	0.16± 0.12	0.15± 0.11
Ours	**0.82± 0.06**	**0.79± 0.05**	**0.85± 0.04**	**0.85± 0.04**

to represent the direction of change in $\mathcal{W}$ space when we change u_i from a to b. The following two properties of $\widehat{\Delta w}(i,a,b)$ are observed:

Property 1. For the i-th variable u_i, $i \in 1, 2, 3, 4$, with any given starting value a and ending value b, we have:

$$\widehat{\Delta w}(i,a,b) \textit{ is almost constant for } \forall u_{\{j\}}.$$

Property 2. For the i-th variable u_i, $i \in 1, 2, 3, 4$, with any given offset vector $\triangle$, we have:

$$\widehat{\Delta w}(i,a,a+\triangle) \textit{ is almost constant for } \forall u_{\{j\}} \textit{and } \forall a.$$

Property 1 states that if the starting and ending values of a certain factor in λ space are fixed, then the direction of change in $\mathcal{W}$ space is stable regardless of the choice of all other factors. Property 2 further indicates that it is unnecessary to fix the starting and ending values – the direction of change in $\mathcal{W}$ space is only decided by the difference between them.

To empirically examine Property 1, we randomly sampled 50 pairs of (a, b) values for each u_i and 100 remaining factors for each pair. For each (a, b) pair, we calculate 100 $\Delta w = w_2 - w_1$ and get 100×100 pairwise cosine distances. We average all these distances for each (a, b) pair, and finally compute the mean and standard derivation of the 50 average distance values from all 50 pairs. Similarly, we examine Property 2 by randomly generating offsets for u_i, and all the results are presented in Table 2. It can be seen that all the cosine similarities are close to 1, indicating the high consistency of $\mathcal{W}$-space direction change. For reference, in the table we also present the statistics obtained using a model trained with the same pipeline but without our imitative-contrastive losses.

5.2. Real Image Embedding and Editing

Based on the above analysis, we show that our method can be used to embed real images into the latent space and edit the factors in a disentangled manner. We present the experimental results on various factors. More results can be found in the *suppl. material* due to space limitation.

A natural latent space for image embedding and editing is the λ space. However, embedding an image to it leads to

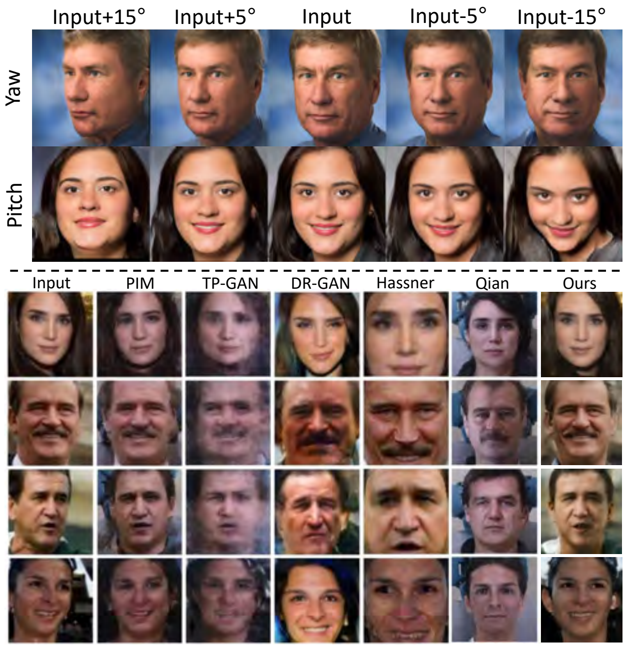

Figure 7: Real image pose manipulation results. **Top**: Precise manipulation of pose angles. **Bottom**: Face frontalization results compared with PIM [53], TP-GAN [21], DR-GAN [47], Hassner *et al*. [16], and Qian *et al*. [37] on LFW. Results of other methods are from [37].

poor image reconstruction. Even inverting to the $\mathcal{W}$ space is problematic – the image details are lost as shown in previous works [1, 41]. For higher fidelity, we embed the image into a latent code w^+ in the $\mathcal{W}+$ space suggested by [1] which is an extended $\mathcal{W}$ space. An optimization-based embedding method is used similar to [1]. However, $\mathcal{W}$ or $\mathcal{W}+$ space is not geometrically interpretable thus cannot be directly used for controllable generation. Fortunately though, thanks to the nice properties of the learned $\mathcal{W}$ space (see Section 5.1), we have the following latent representation editing and image generation method:

$$\begin{aligned} w_s^+ &= G_{syn}^{-1}(x_s) \\ x_t &= G_{syn}(w_s^+ + \Delta w(i,a,b)) \end{aligned} \tag{11}$$

where x_s is an input image and x_t is the targeted image after editing. G_{syn} is the synthesis sub-network of StyleGAN (after the 8-layer MLP). $\Delta w(i,a,b)$ denotes the offset of w induced by changing u_i, the i-th λ-space latent variable, from a to b (see Eq. 10). It can be computed with any $u_{\{j\}}$ (we simply use the embedded one). Editing can be achieved by flexibly setting a and b.

Pose Editing. Figure 7 (top) shows the typical results of pose manipulation where we freely rotate the input face by desired angles. We also test our method with the task of face frontalization, and compare with previous methods. Figure 7 (bottom) shows the results on face images from the LFW dataset [20]. Our method well-preserved the identity-bearing characteristics as well as other contextual information such as hair structure and illumination.

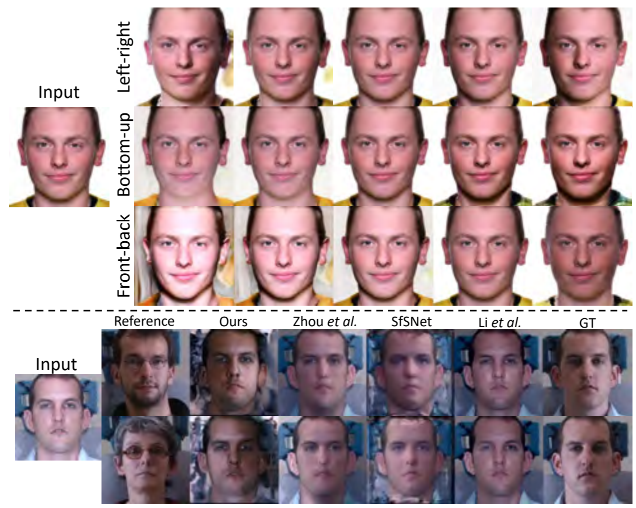

Figure 8: Real image relighting results. **Top**: Light editing for a real image. **Bottom**: Results on a challenging lighting transfer task compared with Zhou *et al*. [54], SfSNet [40], and Li *et al*. [27]. Results of other methods are from [54].

Image Relighting. Figure 8 (top) shows an example of image relighting with our method, where we freely vary the lighting direction and intensity. In addition, we follow the previous methods to evaluate our method on the MultiPIE [15] images. Figure 8 (bottom) shows a challenging case for lighting transfer. Despite the extreme indoor lighting may be outside of the training data, our method still produces reasonable results with lighting directions well conforming with the references.

6. Conclusion and Future Work

We presented ***DiscoFaceGAN*** for disentangled and controllable latent representations for face image generation. The core idea is to incorporate 3D priors into the adversarial learning framework and train the network to imitate the rendered 3D faces. Influence of the domain gap between rendered faces and real images is properly handled by introducing the contrastive losses which explicitly enforce disentanglement. Extensive experiments on disentangled virtual face image synthesis and face image embedding have demonstrated the efficacy of our proposed imitation-contrastive learning scheme.

The generated virtual identity face images with accurately controlled properties could be used for a wide range of vision and graphics applications which we will explore in our future work. It is also possible to apply our method for forgery image detection and anti-spoofing by analyzing real and faked images in the disentangled space.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *IEEE International Conference on Computer Vision*, pages 4432–4441, 2019. 7, 8

[2] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. CVAE-GAN: fine-grained image generation through asymmetric training. In *IEEE International Conference on Computer Vision*, pages 2745–2754, 2017. 2

[3] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Towards open-set identity preserving face synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6713–6722, 2018. 2

[4] Volker Blanz, Thomas Vetter, et al. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, volume 99, pages 187–194, 1999. 2, 3

[5] Tian Qi Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, pages 2610–2620, 2018. 2, 6

[6] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2172–2180, 2016. 1, 2

[7] Bin Dai and David Wipf. Diagnosing and enhancing vae models. *arXiv preprint arXiv:1903.05789*, 2019. 2, 5

[8] Jiankang Deng, Shiyang Cheng, Niannan Xue, Yuxiang Zhou, and Stefanos Zafeiriou. UV-GAN: Adversarial facial uv map completion for pose-invariant face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7093–7102, 2018. 2

[9] 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 *IEEE Computer Vision and Pattern Recognition Workshop on Analysis and Modeling of Faces and Gestures*, 2019. 3, 6

[10] Guillaume Desjardins, Aaron Courville, and Yoshua Bengio. Disentangling factors of variation via generative entangling. *arXiv preprint arXiv:1210.5474*, 2012. 2

[11] Chris Donahue, Zachary C Lipton, Akshay Balsubramani, and Julian McAuley. Semantically decomposing the latent spaces of generative adversarial networks. In *International Conference on Learning Representations*, 2018. 2

[12] Baris Gecer, Binod Bhattarai, Josef Kittler, and Tae-Kyun Kim. Semi-supervised adversarial learning to generate photorealistic face images of new identities from 3d morphable model. In *European Conference on Computer Vision*, pages 217–234, 2018. 2

[13] Zhenglin Geng, Chen Cao, and Sergey Tulyakov. 3d guided fine-grained face manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9821–9830, 2019. 2, 4

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014. 1

[15] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, pages 807–813, 2010. 8

[16] Tal Hassner, Shai Harel, Eran Paz, and Roee Enbar. Effective face frontalization in unconstrained images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4295–4304, 2015. 8

[17] 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. In *Advances in neural information processing systems*, pages 6626–6637, 2017. 7

[18] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. 2

[19] Qiyang Hu, Attila Szabó, Tiziano Portenier, Paolo Favaro, and Matthias Zwicker. Disentangling factors of variation by mixing them. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3399–3407, 2018. 2

[20] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. 2008. 8

[21] 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 *IEEE International Conference on Computer Vision*, pages 2439–2448, 2017. 8

[22] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 7

[23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 1, 2, 5, 7

[24] Hyeongwoo Kim, Pablo Carrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *ACM Transactions on Graphics*, 37(4):163, 2018. 2

[25] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, 2018. 2, 6

[26] Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. In *Advances in Neural Information Processing Systems*, pages 2539–2547, 2015. 2

[27] Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, and Jan Kautz. A closed-form solution to photorealistic image stylization. In *European Conference on Computer Vision (ECCV)*, pages 453–468, 2018. 8

[28] Jinpeng Lin, Hao Yang, Dong Chen, Ming Zeng, Fang Wen, and Lu Yuan. Face parsing with roi tanh-warping. In *IEEE*

Conference on Computer Vision and Pattern Recognition, pages 5654–5663, 2019. 4

[29] Zinan Lin, Kiran Koshy Thekumparampil, Giulia Fanti, and Sewoong Oh. Infogan-cr: Disentangling generative adversarial networks with contrastive regularizers. *arXiv preprint arXiv:1906.06034*, 2019. 2

[30] 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*, volume 97, pages 4114–4124, 2019. 2

[31] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2

[32] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. HoloGAN: Unsupervised learning of 3d representations from natural images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[33] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301, 2009. 2, 3, 4

[34] Guim Perarnau, Joost Van de Weijer, Bogdan Raducanu, and Jose M. Alvarez. Invertible conditional gans for image editing. In *Advances in Neural Information Processing Systems Workshop on Adversarial Training*, 2016. 2

[35] Jingtan Piao, Chen Qian, and Hongsheng Li. Semi-supervised monocular 3d face reconstruction with end-to-end shape-preserved domain transfer. In *IEEE International Conference on Computer Vision*, pages 9398–9407, 2019. 2

[36] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *European Conference on Computer Vision*, pages 818–833, 2018. 2

[37] Yichen Qian, Weihong Deng, and Jiani Hu. Unsupervised face normalization with extreme pose and expression in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9851–9858, 2019. 4, 8

[38] Ravi Ramamoorthi and Pat Hanrahan. An efficient representation for irradiance environment maps. In *SIGGRAPH*, pages 497–500, 2001. 3

[39] Scott Reed, Kihyuk Sohn, Yuting Zhang, and Honglak Lee. Learning to disentangle factors of variation with manifold interaction. In *International Conference on Machine Learning*, pages 1431–1439, 2014. 2

[40] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. Sfsnet: Learning shape, reflectance and illuminance of facesin the wild'. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6296–6305, 2018. 8

[41] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. *arXiv preprint arXiv:1907.10786*, 2019. 7, 8

[42] Yujun Shen, Ping Luo, Junjie Yan, Xiaogang Wang, and Xiaoou Tang. FaceID-GAN: Learning a symmetry three-player gan for identity-preserving face synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 821–830, 2018. 2

[43] Yujun Shen, Bolei Zhou, Ping Luo, and Xiaoou Tang. Facefeat-gan: a two-stage approach for identity-preserving face synthesis. *arXiv preprint arXiv:1812.01288*, 2018. 2

[44] Lingxiao Song, Zhihe Lu, Ran He, Zhenan Sun, and Tieniu Tan. Geometry guided adversarial facial expression synthesis. In *ACM Multimedia Conference on Multimedia Conference*, pages 627–635, 2018. 2

[45] Tiancheng Sun, Jonathan T Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul Debevec, and Ravi Ramamoorthi. Single image portrait relighting. *ACM Transactions on Graphics*, 38(4):79, 2019. 2

[46] Joshua B Tenenbaum and William T Freeman. Separating style and content with bilinear models. *Neural computation*, 12(6):1247–1283, 2000. 2

[47] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1415–1424, 2017. 2, 8

[48] William F Whitney, Michael Chang, Tejas Kulkarni, and Joshua B Tenenbaum. Understanding visual concepts with continuation learning. *arXiv preprint arXiv:1602.06822*, 2016. 2

[49] Taihong Xiao, Jiapeng Hong, and Jinwen Ma. Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. In *European Conference on Computer Vision*, pages 168–184, 2018. 2

[50] Sicheng Xu, Jiaolong Yang, Dong Chen, Fang Wen, Yu Deng, Yunde Jia, and Xin Tong. Deep 3d portrait from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[51] Jiaolong Yang, Peiran Ren, Dongqing Zhang, Dong Chen, Fang Wen, Hongdong Li, and Gang Hua. Neural aggregation network for video face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4362–4371, 2017. 3, 6

[52] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Towards large-pose face frontalization in the wild. In *IEEE International Conference on Computer Vision*, pages 3990–3999, 2017. 2

[53] 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 *IEEE conference on computer vision and pattern recognition*, pages 2207–2216, 2018. 8

[54] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W Jacobs. Deep single-image portrait relighting. In *IEEE International Conference on Computer Vision*, pages 7194–7202, 2019. 2, 8

Disentangled and Controllable Face Image Generation via 3D Imitative-Contrastive Learning
(Supplementary Material)

Yu Deng*[1,2] Jiaolong Yang[2] Dong Chen[2] Fang Wen[2] Xin Tong[2]
[1]Tsinghua University [2]Microsoft Research Asia
{t-yudeng,jiaoyan,doch,fangwen,xtong}@microsoft.com

I. More Implementation Details

VAE structure. We use the same VAE structure for identity α, expression β, illumination γ, and pose θ in λ space. They have three hidden layers for both encoder and decoder. Dimensions of hidden layers in each VAE are 512, 256, 128, and 32 respectively. We use ReLU as the activation layer.

Latent variable dimensions. The z-space variable dimensions are empirically set to 128, 32, 16, and 3 for z_1 to z_4, respectively, and the dimensions of corresponding latent variable in λ-space are 160, 64, 27, and 3. The dimension of the additional noise z_5 is 32.

Training details. We train the λ-space VAEs following the schedule of [1] where we only adopt the first stage. For StyleGAN [3], we follow the standard training procedure of the original method on the FFHQ dataset except that we 1) remove the normalization operation for input latent variable layer, 2) discard the style-mixing strategy, and 3) train up to image resolution of 256×256 due to time constraint.

The StyleGAN training uses a progressive growing strategy [2] where the image resolution gradually increases. It is difficult to directly apply our imitative losses when the resolution is very small and image quality is poor. So when the resolution $\leq 32 \times 32$, we simply use an average l_1 pixel loss between the face regions of generated images and rendered ones with its weight set to 20. The adversarial loss weight is set to 1 throughout the training process. When the resolution grows to 64×64, we discard the pixel loss and apply our imitative losses described in the main paper. We train the network until seeing $15M$ real images to obtain reasonable imitation, with loss weights set as $w_{l_I^{id}} = 3$, $w_{l_I^{lm}} = 500$, $w_{l_I^{sh}} = 10$, and $w_{l_I^{cl}} = 20$. Then we add the contrastive losses and train the network up to seeing $20M$ real images with loss weights set as $w_{l_C^{ex}} = 10$, $w_{l_C^{il_1}} = 10$, and $w_{l_C^{il_2}} = 20$. The balancing weight ω in $l_C^{il_1}$ is set to 1000. During this period, the imitative loss weight $w_{l_I^{lm}}$ is reduced to 100 and others remain unchanged. Note that these loss weights and other hyper-parameters are not carefully tuned.

*This work was done when Yu Deng was an intern at MSRA.

II. More Generation Results

In Figure V and Figure VI, we show more generation results of *DiscoFaceGAN*. Similar to results presented in the main paper, we are able to randomly generate face images with a large variant of identities with diverse poses, illumi-nation conditions and facial expressions. The variations of identity, expression, pose and illumination are highly disen-tangled with each other. Precisely control can be achieved for expression, illumination and pose using the parametric model coefficients.

III. Latent Space Interpolation

In Figure VII, we show some results of latent space interpolation. Since our model learns a disentangled latent space, we can interpolate each factor in the λ space independently. When a certain factor is changing, the corresponding attributes in generated images are changing continuously and smoothly, while attributes related to other factors remain unchanged.

IV. Attribute-Preserving Truncation Trick

In StyleGAN [3], a truncation trick is used to improve the generation quality of the model. Given a latent code w in $\mathcal{W}$ space, we can move it towards a center by $w' = w + (1-\psi)(\bar{w} - w)$, where $\bar{w}$ is the empirical average center in the $\mathcal{W}$ space, and $\psi < 1$. However, naively applying the truncation trick may change all image attributes controlled by the factors in the λ space, whereas we hope to improve the generation quality of the identities while keeping expression, illumination and pose unchanged. Therefore, we propose an attribute-preserving truncation trick based on the latent space properties described in Section 5.1 of the main paper. Specifically, we compute an empiri-

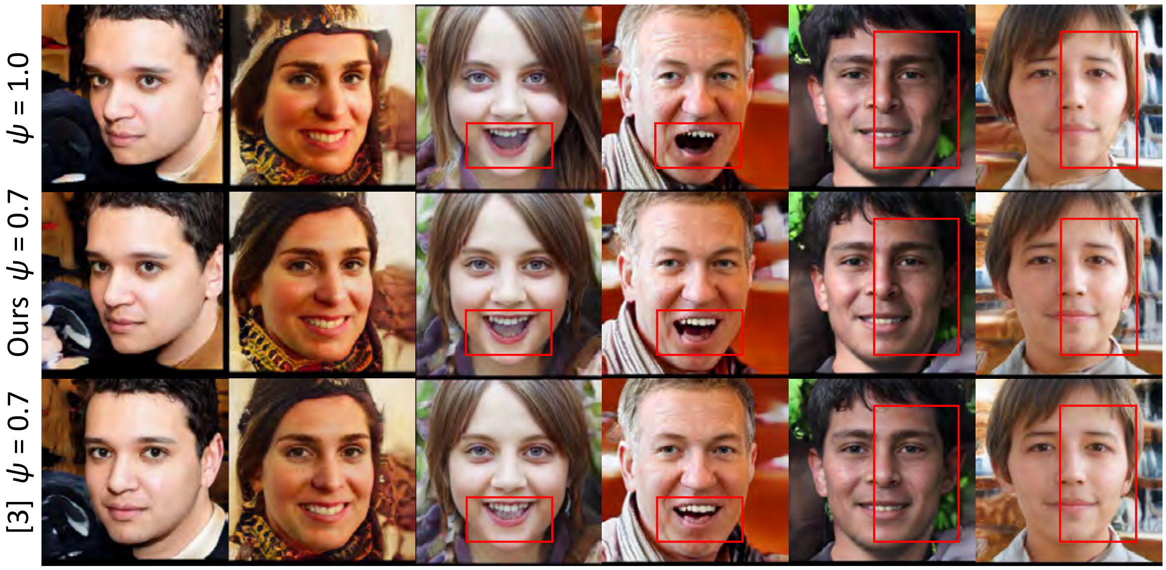

Figure I: Our attribute-preserving truncation trick improves the generation quality meanwhile maintains the pose (first two columns), expression (middle two columns), and illumination (last two columns) of the generated images. The original truncation trick in [3] cannot preserve these attributes.

Figure II: Real image expression transfer results.

cal center $\bar{w}_\alpha = \mathbb{E}_\alpha[w(\alpha, u_{\{j\}} = 0)]$ in the $\mathcal{W}$ space with expression, illumination, and pose set to 0. Then, given a latent code $w(\alpha = a, u_{\{j\}})$, we change it with $w' = w(\alpha = a, u_{\{j\}}) + (1-\psi)(\bar{w}_\alpha - w(\alpha = a, u_{\{j\}} = 0))$. Figure I compares the original truncation trick and ours.

V. Real Image Editing

In Section 5.2 of the main paper, we have presented some results of pose and lighting modification of real images. In Fig. II, we further show some typical results from our method in an expression transfer task. As can be seen, our method successfully transfers the desired expressions to different subjects under various poses and lighting conditions.

VI. Analysis of Image Generation

Since we can flexibly control the generation with disentangled factor variations, we use our method to analyze image generation process of StyleGAN. We provide a stage-

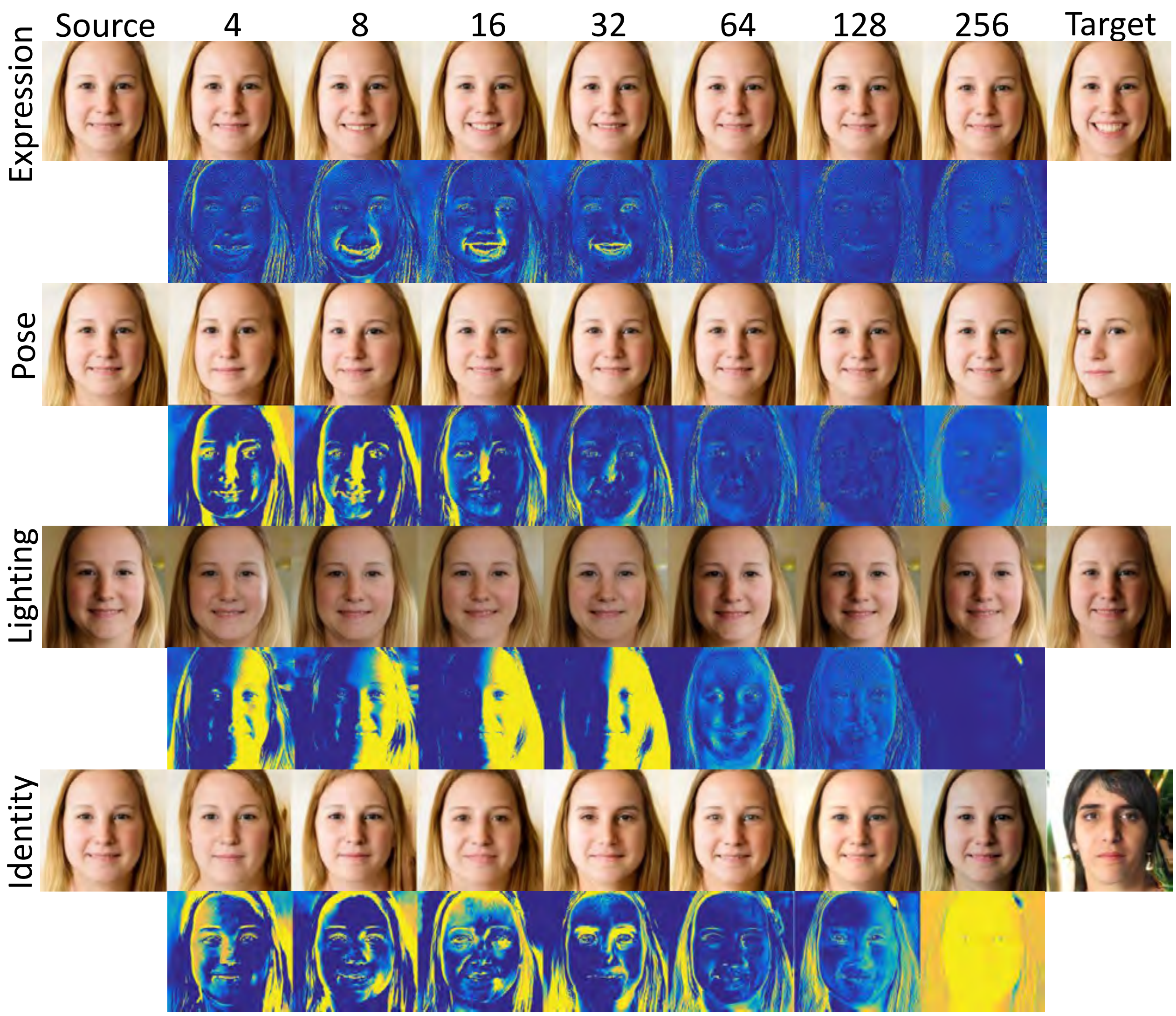

Figure III: Generation results of a 256×256-resolution StyleGAN when changing the parameters of AdaIN layers for each stage. The heatmaps show the color difference between the generated images and the original source image.

by-stage visualization of the impacts on pose, expression, lighting and identity generation. Given an image x_s generated by λ_s, we replace the corresponding w_s vector in the $\mathcal{W}$ space with w_t for the two AdaIN layers at each generation stage (spatial resolution), where w_t comes from another λ_t which differs from λ_s at one factor. Changes on the generated image therefore reflect the impact of each stage on the factor of interest, and Figure III shows one example.

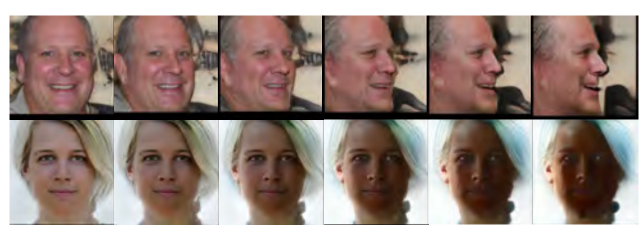

Figure IV: Quality of the generated face images decreases when input factors are out of the distribution of the real image training set.

VII. Limitations

We have demonstrated the effectiveness of our model on disentangled and controllable face image generation. Still, our model has some limitations. Figure IV shows that degraded generation quality of the model under extreme pose and lighting. This is a common out-of-domain issue, resolving which would require using training images with a wider range of distribution beyond FFHQ. In addition, we cannot achieve the control over detailed facial expressions and eye gaze due to the limited ability of 3DMM.

Figure V: More face images generated by our *DiscoFaceGAN*. As shown in the figures, the variations of identity, expression, pose and illumination are highly disentangled, and we can precisely control expression, illumination and pose.

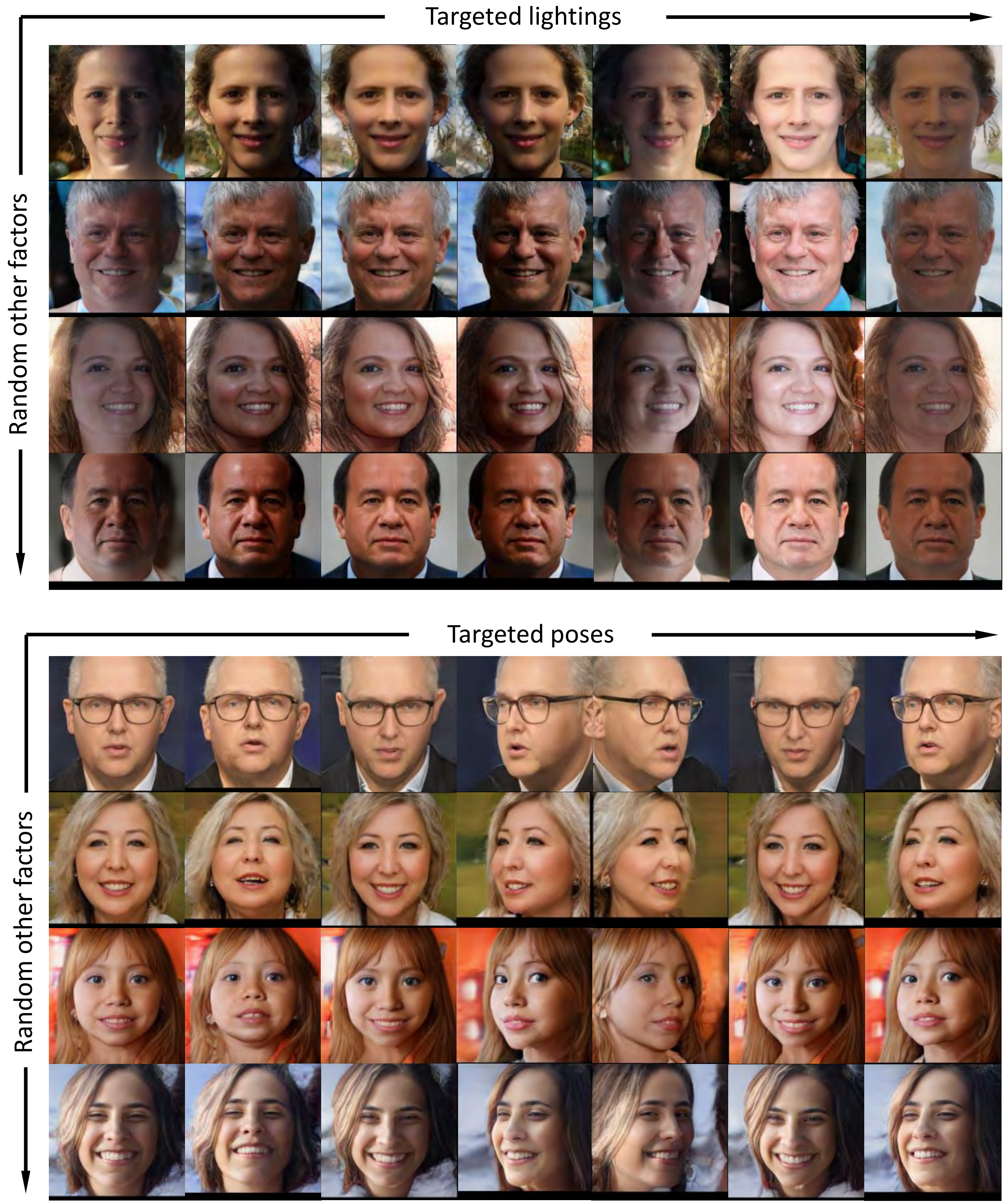

Figure VI: More face images generated by our *DiscoFaceGAN*. As shown in the figures, the variations of identity, expression, pose and illumination are highly disentangled, and we can precisely control expression, illumination and pose.

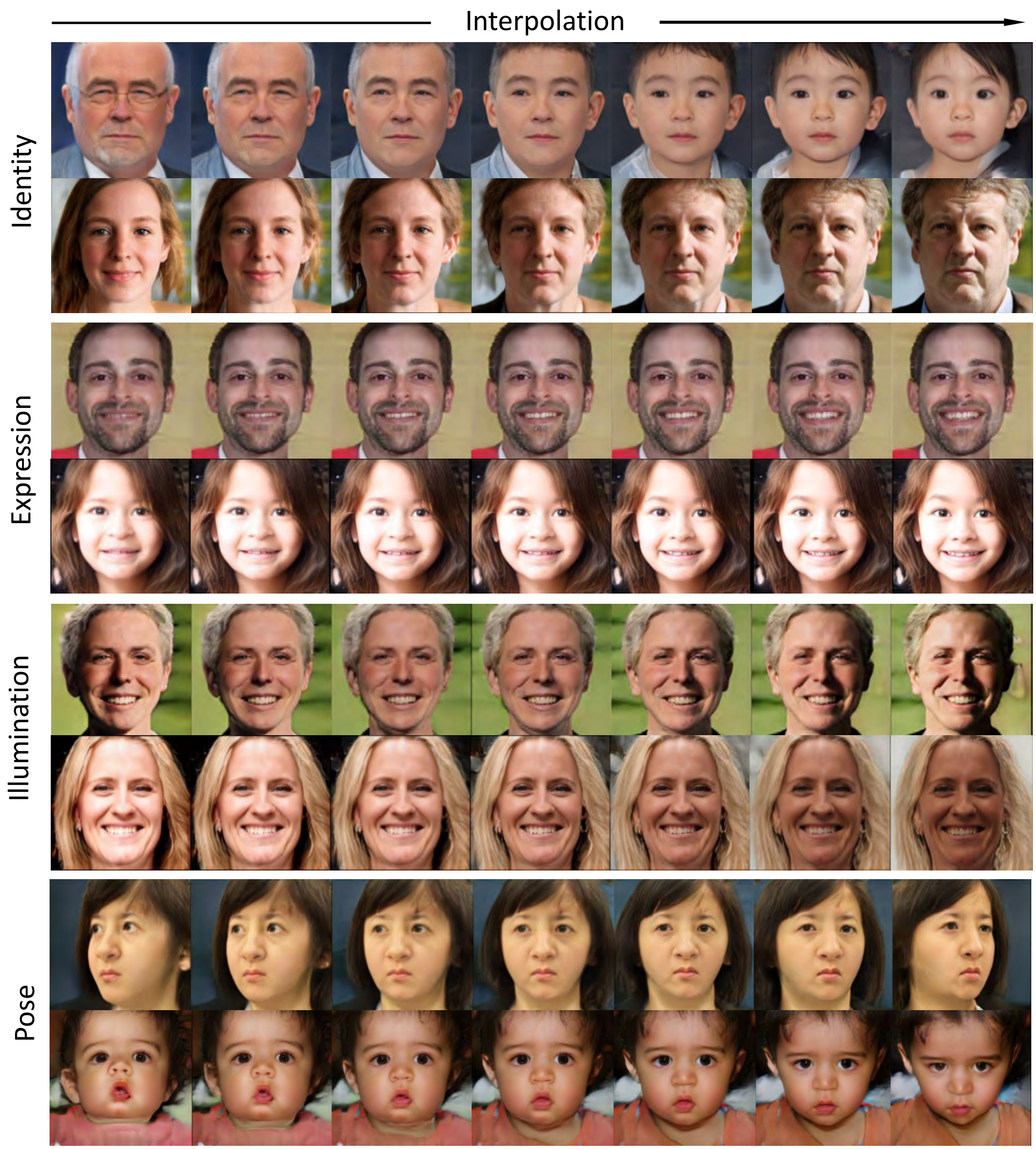

Figure VII: Latent space interpolation result. We can interpolate each factor independently and the corresponding outcome images are reasonable.

References

[1] Bin Dai and David Wipf. Diagnosing and enhancing vae models. *arXiv preprint arXiv:1903.05789*, 2019. 1

[2] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 1

[3] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 1, 2