GAN-Control: Explicitly Controllable GANs

Alon Shoshan Nadav Bhonker Igor Kviatkovsky Gérard Medioni
Amazon
{alonshos, nadavb, kviat, medioni}@amazon.com

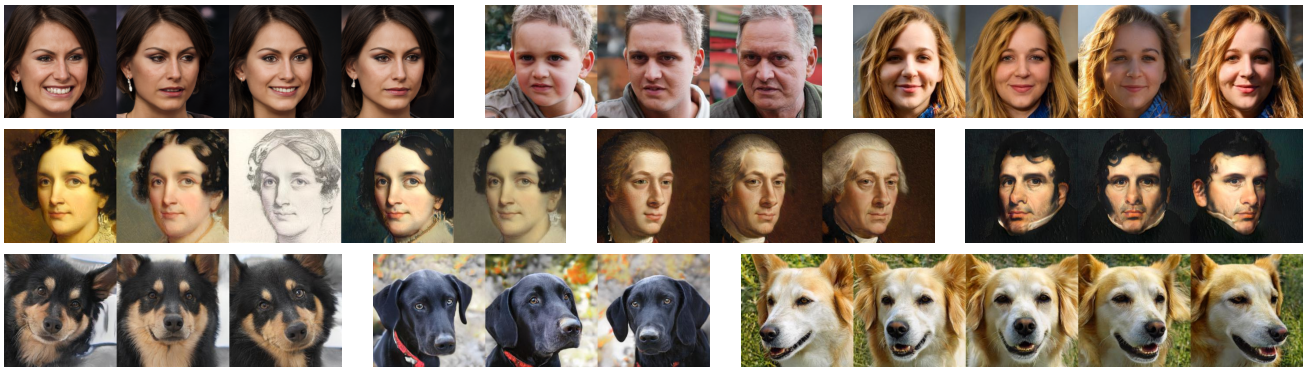

Figure 1: We propose a framework for training GANs in a disentangled manner which allows for explicit control over generation attributes. Our method is applicable to diverse controls in various domains. First row (left to right) demonstrates our control over facial expression, age and illumination of human portraits. Second row (left to right) demonstrates our control over artistic style, age and pose of paintings. Third row demonstrates our pose control over faces of dogs.

Abstract

We present a framework for training GANs with explicit control over generated facial images. We are able to control the generated image by settings exact attributes such as age, pose, expression, etc. Most approaches for manipulating GAN-generated images achieve partial control by leveraging the latent space disentanglement properties, obtained implicitly after standard GAN training. Such methods are able to change the relative intensity of certain attributes, but not explicitly set their values. Recently proposed methods, designed for explicit control over human faces, harness morphable 3D face models (3DMM) to allow fine-grained control capabilities in GANs. Unlike these methods, our control is not constrained to 3DMM parameters and is extendable beyond the domain of human faces. Using contrastive learning, we obtain GANs with an explicitly disentangled latent space. This disentanglement is utilized to train control-encoders mapping human-interpretable inputs to suitable latent vectors, thus allowing explicit control. In the domain of human faces we demonstrate control over identity, age, pose, expression, hair color and illumination. We also demonstrate control capabilities of our framework in the domains of painted portraits and dog image generation. We demonstrate that our approach achieves state-of-the-art performance both qualitatively and quantitatively.

1. Introduction

Generating controllable photorealistic images has applications spanning a variety of fields such as cinematography, graphic design, video games, medical imaging, virtual communication and ML research. For faces in particular, impressive breakthroughs were made. As an example, in the film industry, computer generated characters are replacing live actor footage. Earlier work on controlled face generation primarily relied on 3D face rig modeling [32, 43], controlled by 3D morphable face model parameters, such as 3DMM [9, 19]. While easily controllable, such methods tend to suffer from low photorealism. Other methods that rely on 3D face scanning techniques may provide highly photorealistic images, but at a significant cost and limited variability. Recent works on high resolution images synthesis using generative adversarial networks (GANs) [21] have demonstrated the ability to generate photorealistic faces of novel identities, indistinguishable from those of real humans [27, 29, 30]. However, these methods alone lack in-

terpretability and control over the generative process, compared to the 3D graphic alternatives.

These results have inspired the community to explore ways to benefit from both worlds – generating highly photorealistic faces using GANs while controlling their fine-grained attributes, such as pose, illumination and expression with 3DMM-like parameters. Deng *et al*. [15], Kowalski *et al*. [31] and Tewari *et al*. [50] introduce explicit control over GAN-generated faces, relying on guidance from 3D face generation pipelines. Along with the clear benefits, such as the precise control and perfect ground truth, reliance on such 3D face models introduces new challenges. For example, the need to overcome the synthetic-to-real domain gap [31, 15]. Finally, all these methods' expressive power is bounded by the capabilities of the model they rely on. In particular, it is not possible to control human age if the 3D modeling framework does not support it. It is also impossible to apply the same framework to different but similar domains, such as paintings or animal faces, if these assets are not supported by the modeling framework. All of these stand in the way of creating a simple, generic and extendable solution for explicitly controllable GANs.

In this work we present a unified approach for training a GAN to generate high-quality, controllable images. Specifically, we demonstrate our approach in the domains of facial portrait photos, painted portraits and dogs (see Fig. 1). We depart from the use of the highly detailed 3D face models [15, 31, 50] in favor of supervision signals provided by a set of pre-trained models, each controlling a different feature. We show that our approach significantly simplifies the generation framework, does not compromise image quality or control accuracy, and allows us to control additional aspects of facial appearances, which cannot be modeled by graphical pipelines. We achieve this by combining several concepts. We construct the GAN's latent space as a composition of sub-spaces, each corresponding to a specific property. During training, we enforce images generated by identical latent sub-vectors to have similar properties, as predicted by some off-the-shelf model. Respectively, images generated by different latent sub-vectors are enforced to have different predicted properties. As a result, disentanglement between the latent sub-spaces is achieved. Finally, to allow for human-interpretable control, for each attribute we train an encoder converting values from its feasible range to its corresponding sub-latent space. As an additional application, we present a novel image projection approach suitable for disentangled latent spaces.

We summarize our contributions as following:

1. We present a novel state-of-the art approach for training explicitly controllable, high-resolution GANs.

2. Our approach is extendable to attributes beyond those supported by 3D modeling and rendering frameworks, making it applicable to additional domains.

3. We present a disentangled projection method that enables real image editing.

2. Related work

Generative adversarial networks [21] introduced new possibilities to the field of image generation and synthesis. Currently, state-of-the-art GANs [10, 27, 29, 30] can produce high-resolution images that are indistinguishable from real ones. Next, we provide an overview of different approaches to control the generated output of GANs.

Relative control over image generation: A widely studied approach for controlling the generated images of GANs is by exploiting the inherent disentanglement properties of their latent space [26, 53, 22, 44, 7]. Härkönen *et al*. [22] use principal component analysis (PCA) in latent space to identify directions that correspond to image attributes. Shen *et al*. [44] use off-the-shelf binary classifiers to find separation boundaries in the latent space where each side of the boundary corresponds to an opposite semantic attribute (*e.g*., young vs. old). Traversing a latent vector closer to or further from a boundary translates to increasing or decreasing the corresponding attribute intensity. While simple, these methods may exhibit entanglement, *i.e*., changing one attribute affects others. In [18, 45] the above is mitigated by disentangling the GAN's latent space during training. While the above methods allow for relative control over the generation (*e.g*., turn the face older or rotate the face towards the left), they do not provide explicit control (*e.g*., generate a 40 years old face, rotated 30° to the left).

Explicit control over image generation: Conditional GANs [33, 36, 34, 10] have been widely employed to control the generation by incorporating a class label inference loss term. All these works support conditioning on a single discrete (categorical) variable and are not suitable for continues variables, as was broadly discussed in Ding *et al*. [17]. Furthermore, none of the above works address the problem of controlling multiple attributes at once. Recently, three novel methods were proposed to allow fine-grained explicit control over *de novo* face image generation: StyleRig [50], DiscoFaceGAN [15] (DFG), and CONFIG [31]. These methods propose solutions for translating controls of 3D face rendering models to GAN-generating processes. Both StyleRig and DFG utilize 3DMM [9] parameters as controls in the generation framework. This restricts both approaches to provide controls only over the expression, pose and illumination, while preserving identity (ID). CONFIG uses a custom 3D image rendering pipeline to generate an annotated synthetic dataset. This dataset is later used to acquire controls matching the synthetic ground truth, allowing CONFIG to add controls such as hair style and gaze.

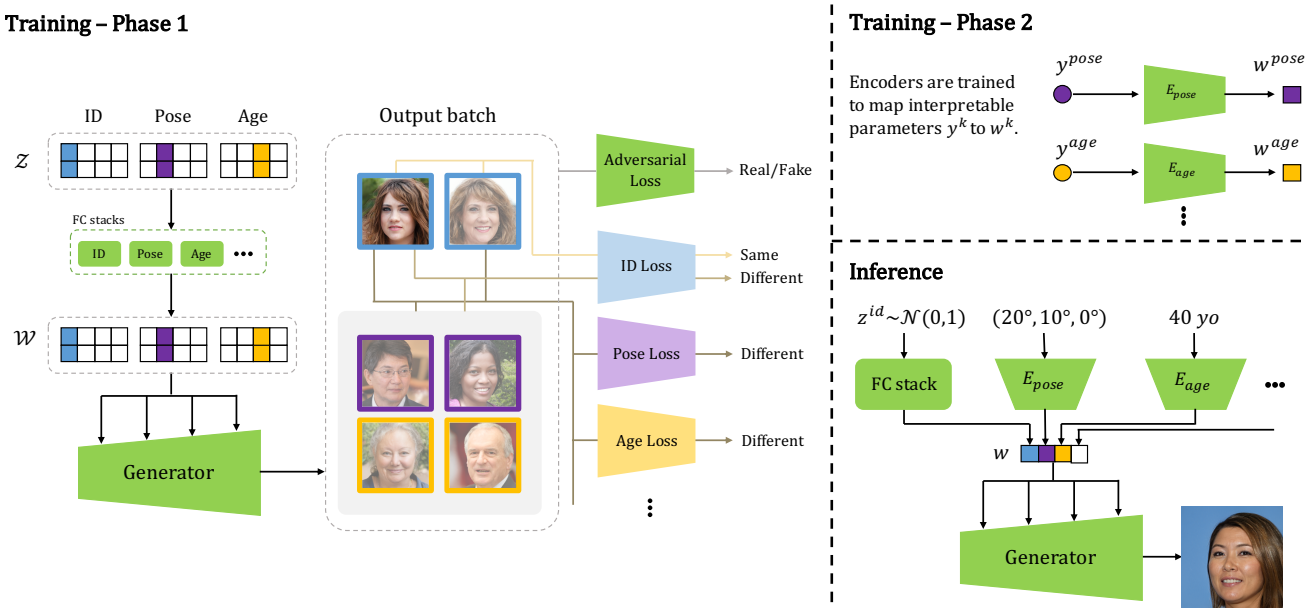

Figure 2: **Explicitly controllable GAN**: In Phase 1, we construct every batch so that for each attribute, there is a pair of latent vectors sharing a corresponding sub-vector, $\mathbf{z}^k$. In addition to the adversarial loss, each image in the batch is compared in a contrastive manner, attribute-by-attribute, to all others, taking into account if it has the same or a different sub-vector. In Phase 2, encoders are trained to map interpretable parameters to suitable latent vectors. Inference: An explicit control over the attribute k is achieved by setting the E_k's input to a required value.

Producing such datasets is hard and requires professional handcrafted 3D assets. We emphasize that these methods are only applicable in the domain of human faces, and only to the controls parametrized by 3D face models. In contrast to the above methods, our approach does not rely on 3D face rendering frameworks. Rather, it relies on our ability to estimate such properties.

Image editing: Rather than generating images *de novo*, these methods receive an image as input and manipulate its attributes either by using image-to-image translation techniques [58, 52, 39, 37, 11, 12, 25], by incorporating pre-trained models to supervise GAN's training [46, 8, 23, 54], or by projecting the image to the GAN's latent space and manipulating it [57, 6, 51, 38, 49, 59, 56]. Our work focuses on controllable *de novo* image generation, but also allows editing real images via projection to latent space.

3. Proposed approach

In this section we present our framework for training explicitly controllable GANs. Our approach is simple yet effective and is comprised of two phases (see Fig. 2):

- **Disentanglement by contrastive learning:** training a GAN with explicitly disentangled properties. As a result, the latent space is divided into sub-spaces, each encoding a different image property.
- **Interpretable explicit control:** for each property, an MLP encoder is trained to map control parameter values to a corresponding latent sub-space. This enables explicit control over each one of the properties.

3.1. Disentanglement by contrastive learning

The approach builds on the StyleGAN2 [30] architecture. Initially, we divide both latent spaces, $\mathcal{Z}$ and $\mathcal{W}$ to $N+1$ separate sub-spaces, $\{\mathcal{Z}^k\}_{k=1}^{N+1}$, and $\{\mathcal{W}^k\}_{k=1}^{N+1}$, where N is the number of control properties. Each sub-space is associated with an attribute (*e.g.*, ID, age etc.) except for the last one. Similarly to Deng *et al.* [15] the last sub-space encodes the rest of the image properties that are not controllable. We modify the StyleGAN2 architecture so that each control has its own 8-layered MLP. We denote $\mathbf{z} = (\mathbf{z}^1\mathbf{z}^2\dots\mathbf{z}^{N+1})$ and $\mathbf{w} = (\mathbf{w}^1\mathbf{w}^2\dots\mathbf{w}^{N+1})$ the concatenation of the sub-vectors in both latent spaces. The combined latent vector, $\mathbf{w}$, is then fed into the generator.

Next, we describe how we enforce disentanglement during training. Let $\mathcal{I}_i = G(\mathbf{z}_i)$ denote an image generated from a latent vector $\mathbf{z}_i$ and let $B = \{\mathbf{z}_i\}_{i=1}^{N_B}$ denote a latent vector batch of size N_B. We define our factorized-contrastive loss as:

$$L_c = \sum_{\substack{\mathbf{z}_i,\mathbf{z}_j \in B \\ i \neq j}} \sum_{k=1}^{N} l_k(\mathbf{z}_i, \mathbf{z}_j), \tag{1}$$

where l_k is a contrastive loss component for attribute k. We

define the per-attribute contrastive loss as,

$$l_k(\mathbf{z}_i, \mathbf{z}_j) = \begin{cases} \frac{1}{C_k^+} \max\left(d_k(\mathcal{I}_i, \mathcal{I}_j) - \tau_k^+, 0\right), & \mathbf{z}_i^k = \mathbf{z}_j^k \\ \frac{1}{C_k^-} \max\left(\tau_k^- - d_k(\mathcal{I}_i, \mathcal{I}_j), 0\right), & \text{otherwise} \end{cases} \quad (2)$$

where $\mathbf{z}_i^k$ denotes the k-th sub-vector of $\mathbf{z}_i$, d_k is the distance function for attribute k, $\tau_k^{\pm}$ are the per-attribute thresholds associated with same and different sub-vectors and $C_k^{\pm}$ are constants that normalize the loss according to the number of same and different loss components, *i.e.* $C_k^+ = \sum_{i,j} \mathbb{1}\{\mathbf{z}_i^k = \mathbf{z}_j^k\}$ and $C_k^- = \sum_{i,j} \mathbb{1}\{\mathbf{z}_i^k \neq \mathbf{z}_j^k\}$.

We construct each training batch to contain pairs of latent vectors that share one sub-vector, *i.e.*, for each attribute, $k \in \{1, \ldots, N\}$, we create a pair of latent vectors, $\mathbf{z}_i$ and $\mathbf{z}_j$, where $\mathbf{z}_i^k = \mathbf{z}_j^k$ and $\mathbf{z}_i^r \neq \mathbf{z}_j^r$ for $r \in \{1, \ldots, N+1\}, r \neq k$. For example, let us assume that the generator has produced a batch of size $N_b > 2$, where images $\mathcal{I}_0$ and $\mathcal{I}_1$ share the same $\mathbf{z}^{ID}$ (see the pair of images with the blue frame in Fig. 2). The ID component of the contrastive loss, l_{ID}, will penalizes the dissimilarities between the $\mathcal{I}_0$'s and $\mathcal{I}_1$'s IDs and the similarities between $\mathcal{I}_0$'s or $\mathcal{I}_1$'s ID to the IDs of all other images in the batch. The other loss components (age, pose, *etc*.) will penalize for similarity between $\mathcal{I}_0$ and any other image in the batch. The losses for all other images in the batch are constructed in the same manner.

To be able to control a specific attribute of the generated image, we assume that we are given access to a differentiable function $M_k : \mathcal{I} \to \mathbb{R}^{D_k}$, mapping an image to a D_k-dimensional space. We assume that the projected images with similar attribute values fall close to each other, and images with different attribute values fall far from one another. Such requirements are met by most neural networks trained with either a classification or a regression loss – for example, a model estimating the head pose or the person's age. We define the k's attribute distance between two images $\mathcal{I}_i$ and $\mathcal{I}_j$ as their distance in the corresponding embedding space:

$$d_k(\mathcal{I}_i, \mathcal{I}_j) = dist(M_k(\mathcal{I}_i), M_k(\mathcal{I}_j)), \quad (3)$$

where $dist(\cdot, \cdot)$ is a distance metric, *e.g.*, L_1, L_2, cosine-distance, *etc*. For example, to capture the ID property, a face recognition model, M_{ID}, is used to extract embedding vectors from the generated images. Then, the distances between the embedding vectors are computed using the cosine-distance.

In Section 4 we demonstrate that as the result of training with this architecture and batch sampling protocol, we achieve disentanglement in the GAN's latent space. While such disentanglement allows to assign a randomly sampled value to each individual attribute, independently of the others, additional work is required for turning such control explicit and human-interpretable, *e.g.*, generate a human face image with a specific user-defined age.

3.2. Interpretable explicit control

We propose a simple procedure to allow explicit control of specific attributes. We train a mapping $E_k : y^k \to \mathbf{w}^k$, where y^k is a human-interpretable representation of the attribute (*e.g.*, age = 20yo, pose = $(20°, 5°, 2°)$, *etc*.). Given a trained disentangled GAN, we train N encoders $\{E_k\}_{k=1}^N$, one for each attribute (see Training-Phase 2 in Fig. 2). Then, at inference time we can synthesize images using any combination of sub-vectors $\{\mathbf{w}^k\}_{k=1}^{N+1}$, where $\mathbf{w}^k$ is either controlled explicitly using E_k or sampled from $\mathbf{z}^k$ and consequently mapped to $\mathbf{w}^k$ (see Inference in Fig. 2).

To train the control encoders, we randomly sample N_s latent vectors $\{\mathbf{z}_i\}_{i=1}^{N_s}$ and map them to the intermediate latent vectors, $\{\mathbf{w}_i\}_{i=1}^{N_s}$. Then, for each attribute, k, we map $\mathbf{z}_i$ to a predicted attribute value $y_i^k = Q_k(M_k(G(\mathbf{z}_i)))$, where $Q_k(M_k(\cdot))$ is equivalent to applying the attribute predictor. Thus we obtain N distinct datasets $\{\{\mathbf{w}_i^k, y_i^k\}_{i=1}^{N_s}\}_{k=1}^N$, where for each intermediate sub-vector $\mathbf{w}^k$ there is a corresponding attribute predicted from the image it produced. We then train N encoders, each on its corresponding dataset. In our experiments we show that despite its simplicity, our encoding scheme does not compromise control accuracy compared to other methods.

4. Experiments

In this section we present experiments on the domain of faces and paintings that demonstrate the flexibility of the proposed approach. Additional experiments for images of dogs are presented in the supplementary. We quantitatively compare our approach to recent published approaches.

4.1. Face generation

Implementation details: We use the FFHQ dataset [29] downsampled to 512x512 resolution. The latent spaces $\mathcal{Z}$ and $\mathcal{W}$ are divided into the following sub-spaces: ID, pose, expression, age, illumination, hair color and "other". Next, we list the models, M_k, that were used to compute the distance measures, d_k, for each one of the attributes. For the ID, head-pose, expression, illumination, age and hair color we used ArcFace [14], Ruiz *et al*. [42], ESR [48], the γ output of R-Net [16], Dex [41], average color of hair segmented by PSPNet [55], respectively (additional details in supplementary material). In the second phase (Section 3.2), we train five encoders ($E_{pose}, E_{exp}, E_{age}, E_{illum}, E_{hair}$), each composed of a 4-layered MLP. The input to our control encoder is defined as follows: $y^{age} \in [15, 75]$ years-old (yo), $y^{pose} \in [-90°, 90°]^3$ is represented by the Euler angles $\theta = \{\text{Pitch}, \text{Yaw}, \text{Roll}\}$, $y^{illum} \in \mathbb{R}^{27}$ is represented by the γ Spherical Harmonics (SH) coefficients approximating scene illumination [40], $y^{exp} \in \mathbb{R}^{64}$ is represented by the β expression coefficients of the 3DMM [9] model, $y^{hair} \in [0, 255]^3$ is represented by the mean RGB values.

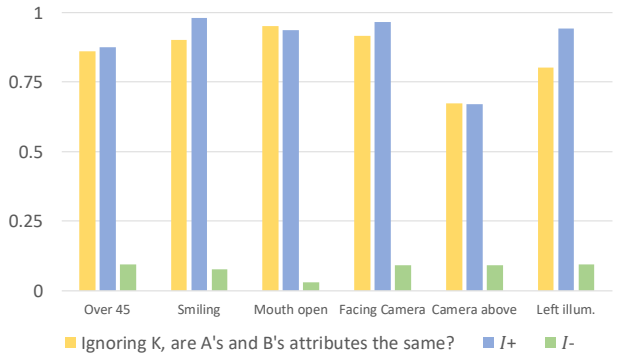

Figure 3: **Disentanglement user study:** Blue and green bars show whether users agree that a given attribute is present or lacking in images ($\mathcal{I}_+$, $\uparrow$), ($\mathcal{I}_-$, $\downarrow$) respectively. The yellow bars measure whether the users agree that the other attributes are maintained.

GAN Version	Ours 512x512	DFG [15] 256x256	CONFIG [31] 256x256
Vanilla	3.32	5.49	33.41
Controlled	5.72	12.9	39.76

Table 1: **FID↓ score for different methods on FFHQ:** second row shows the dataset resolution. Note that the FID scores cannot be compared between columns since every method uses different pre-processing for the FFHQ dataset (*e.g.*, image size, alignment, cropping).

	Ours	DFG	CONFIG
Synthetic comparison	**67%**	22%	11%
Synthetic vs. real	**47%**	27%	16%

Table 2: **Photorealism user studies↑:** (First row) users were asked to vote for the most realistic image from triplets of synthetic images (Ours, DFG, CONFIG). (Second row) users were shown pairs of images – one synthetic and one from the FFHQ dataset – and were asked to choose the real one from the two.

Photorealism: Table 1 shows FID [24] scores of StyleFaceGAN [15], CONFIG [31] (image resolution 256x256) and our approach (image resolution 512x512). The table also shows the FID of the corresponding baseline GANs: StyleGAN [29], HoloGAN [35] and StyleGAN2 [30]. Our FID score is calculated without the use of the truncation trick [10, 29]. For DFG and CONFIG the FID score is taken from the corresponding papers. Similarly to the other works, we observe a deterioration in FID when control is introduced. However, due to the different image resolutions and data pre-processing steps, the numbers are not directly comparable. To make a clear image quality comparison between all three methods, we conducted two photorealism user studies, using Amazon Mechanical Turk. In the first, users were shown 1K triplets of synthetic images, one from each method in a random order. Users were asked to vote for the most realistic image of the three. Each triplet was evaluated by three participants. In the second study, users were shown 999 pairs of images. Each pair contains one real image from the FFHQ dataset and an image generated by one of the three methods. For each method, 333 image pairs were evaluated by three different users. All the synthetic images in this experiment were generated using the truncation trick with $\Psi = 0.7$ (Ours and DFG use the attribute-preserving truncation trick [15]), and all images were resized to 256x256 resolution. From Table 2 it is evident that our method achieves the highest photorealism. Surprisingly, our method reaches a near perfect result of 47% when compared to FFHQ, *i.e.*, users were barely able to distinguish between our images and the ones from FFHQ in terms of photorealism. We note that differences in image quality may depend on the base model that was used (HoloGAN, StyleGAN, StyleGAN2).

Control	Ours	DFG	CONFIG	FFHQ
Pose [°]	$2.29_{\pm 1.31}$	$3.92_{\pm 2.1}$	$6.9_{\pm 4.7}$	$23.8_{\pm 14.6}$
Age [yo]	$2.02_{\pm 1.38}$	N/A	N/A	$16.95_{\pm 12.9}$
Exp.	$3.68_{\pm 0.7}$	$4.07_{\pm 0.7}$	N/A[a]	$4.45_{\pm 0.9}$
Illum.	$0.32_{\pm 0.13}$	$0.29_{\pm 0.1}$	N/A[a]	$0.62_{\pm 0.2}$
Hair color	$0.13_{\pm 0.18}$	N/A	N/A[a]	$0.34_{\pm 0.25}$

Table 3: **Control precision↓:** Comparison of average distance between input controls to resulted image attribute. Last column shows the average distance between random samples in the FFHQ dataset.

[a]CONFIG uses different controls for expression illumination and hair color.

ID	Ours	Ours$_{+age}$	DFG	CONFIG
Same↓	$\mathbf{0.68_{\pm 0.19}}$	$0.75_{\pm 0.2}$	$0.83_{\pm 0.3}$	$1.07_{\pm 0.29}$
Not same↑	$\mathbf{1.9_{\pm 0.24}}$	$\mathbf{1.9_{\pm 0.24}}$	$1.73_{\pm 0.24}$	$1.63_{\pm 0.25}$

Table 4: **Identity preservation:** First row shows the mean embedding distance between generated images with the same $\mathbf{z}^{ID}$ (in Ours$_{+age}$, $\mathbf{z}^{age}$ is also changed). Second row shows the mean embedding distance between randomly generated images. For comparison, the mean embedding distance between 10K FFHQ images is 1.89 ± 0.21.

Explicit control analysis: To validate that we indeed have an explicit control over the output of our model, we perform a control precision comparison. 10K images are randomly chosen from FFHQ and their attributes are predicted to produce a pool of feasible attributes that appear in real images. For each attribute in the pool, y_i^k, we generate a corresponding image. Then, we predict the attribute

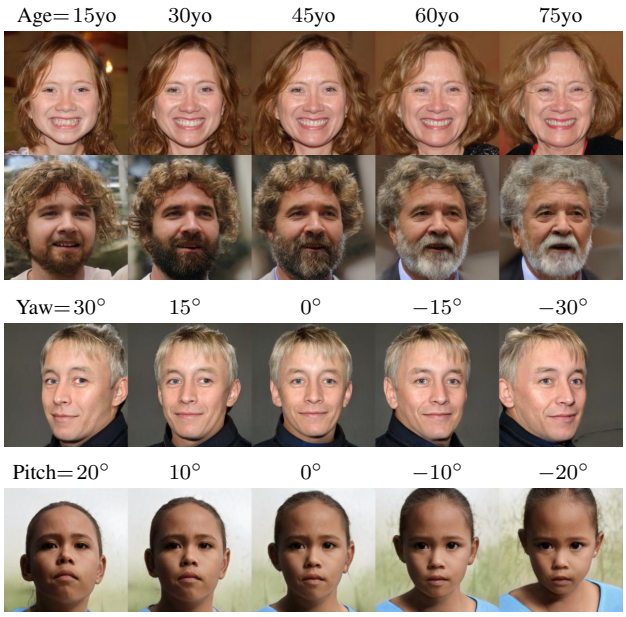

Figure 4: **Controlling age and pose:** Rows 1-2 show generation results using E_{age}. Rows 3-4 show generation results using E_{pose}.

value from the generated image, $\hat{y}_i^k$, and measure the Euclidean distance between the two. More details are provided in the supplementary material. Table 3 shows the comparison of the control precision between the methods. The results demonstrate that we can achieve explicit control of the attributes that is comparable or better than other methods.

ID preservation analysis: We use ArcFace [14] to extract embedding vectors of generated images to compare identity preservation to other methods. This is done by generating 10K image pairs that share the ID attribute and have different pose, illumination and expression attributes. We choose to modify these as they are common to all three methods. To demonstrate the ability of our method to preserve the ID even at different ages, we report results for Ours$_{+age}$ where each image in a pair is generated using a different $\mathbf{z}^{age}$ vector. The results in Table 4 demonstrate that our method achieves the highest identity preservation.

Disentanglement user study: We conducted a user study similar to the one reported in CONFIG [31]. For each attribute, k, we generate a pair of images, $\mathcal{I}_+, \mathcal{I}_-$. The attribute for $\mathcal{I}_+$ is set to y_+^k (*e.g.*, smiling face) and the attribute for $\mathcal{I}_-$ is set to a semantically opposite value y_-^k (*e.g.*, sad face). Users are then asked to evaluate the presence of y_+^k in $\mathcal{I}_+$ and $\mathcal{I}_-$ on a 5-level scale. In addition, for every pair of images the users are asked to evaluate to what extent all other attributes, apart from k, are preserved. In total, 50 users have evaluated 1300 pairs of images. Fig. 3 clearly shows that the attributes of the generated images are perceived as disentangled.

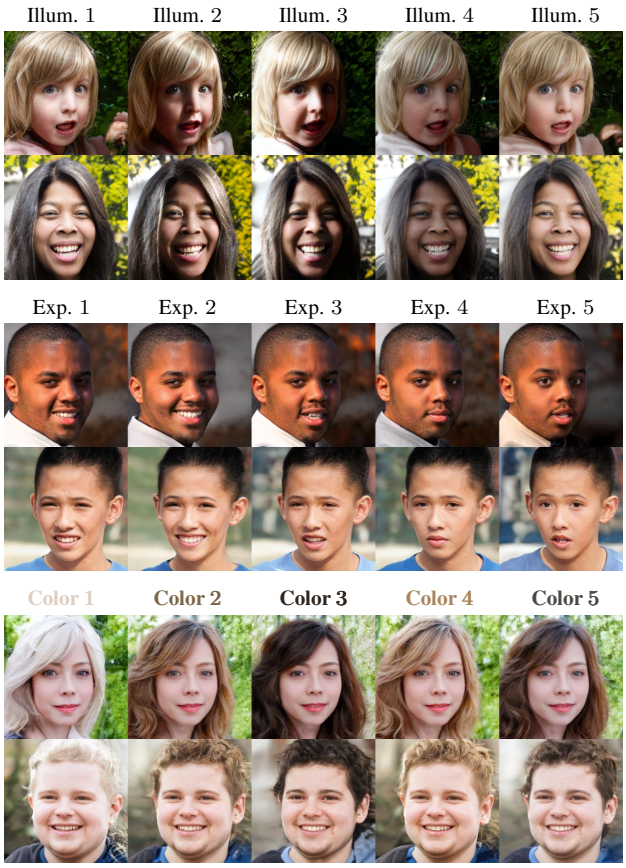

Figure 5: **Controlling illumination, expression and hair color:** Rows 1-2 show generation results using E_{illum}. Rows 3-4 show generation results using E_{exp}. Rows 5-6 show generation results using E_{hair}. Each column has the same attribute matching the control input.

Qualitative evaluation: Next we show editing results of generated images via the control encoders E_k. Fig. 4 shows explicit control over age and pose of faces using E_{age} and E_{pose}. Interestingly, as the age is increased the model tends to generate glasses as well as more formal clothing. Two other prominent features are graying of the hair and the addition of wrinkles. Fig. 5 shows control over illumination and expression using E_{illum} and E_{exp}.

4.2. Painting generation

Implementation details: We use MetFaces [28], 1,336 images downsampled to 512x512 resolution. In addition to the traditional StyleGAN2 and our explicit disentanglement training schemes, we use the method of non-leaking augmentation by Karras *et al*. [28] for training GANs with limited data. We use the same M_k models as in our face generation scheme with the following modifications: (1) the illumination and hair color controls are removed, (2) a control for image style is added. The style similarity distance d_{style}

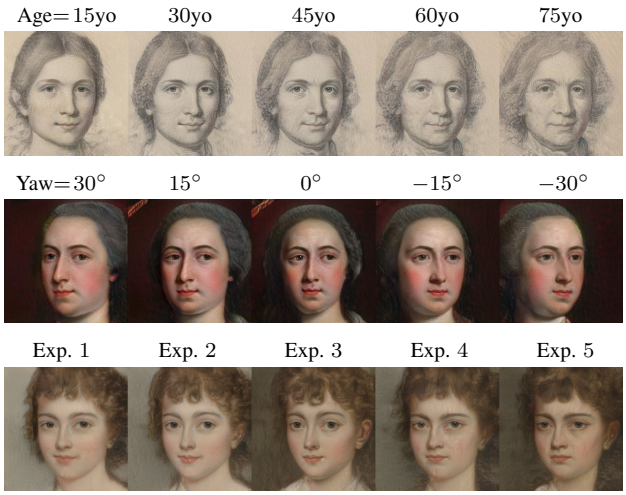

Figure 6: **Control of paintings:** Generation results using E_{age}, E_{pose} and E_{exp}.

Figure 7: **Artistic style for paintings:** We can change the $\mathbf{z}^{style}$ latent to produce same portraits with different style.

is computed similarly to the style loss introduced for style transfer by Gatys *et al*. [20] where M_{style} is a VGG16 [47] network pre-trained on ImageNet [13].

Photorealism: The FID scores are 28.58 and 26.6 for our controlled and for the baseline models, respectively.

Qualitative evaluation: Fig. 6 shows our control over age, pose and expression using E_{age}, E_{pose} and E_{exp}. Note that the expression control for this task is rather limited. We suspect this is due to the low variety of expressions in the dataset. The control over these attributes demonstrates that the control networks do not necessarily need to be trained on the same domain on which the GAN is being trained, and that some domain gap is tolerable. Fig. 7 shows that our method can also disentangle artistic style allowing to change the style without affecting the rest of the attributes.

4.3. Ablation study

In this section we explore two alternative approaches to our framework. (1) Training the GAN end-to-end in a single training phase. In every iteration, the inputs to the model are control attribute values, (y^k), rather than latent vectors. We use the pre-trained models (the same ones as in our two-phase approach) to penalize for disagreement between the attribute values, predicted for each generated image, and the input controls (attribute matching loss). For a fair comparison to our approach, we avoid the harder task of mapping an ID embedding to an image, by maintaining the ID contrastive terms as in Sec. 3.1. We use two configurations of matching loss coefficients where for the first model (E2E) the coefficients are 10 times smaller in magnitude than for the second one (E2E-10x). (2) Instead of training a disentangled GAN in Phase 1, we train attribute encoders for a pre-trained StyleGAN2 (NoDis). Since StyleGAN2's $\mathcal{W}$ space is not divided into disentangled sub-spaces, we train a single encoder mapping all inputs (together) to $\mathcal{W}$. Further implementation details of alternatives 1 and 2 are provided in the supplementary.

In Table 5 we compare our two-phased approach to both alternatives using the control precision, ID preservation and

	Ours	E2E	E2E-10x	NoDis
Control precision ↓				
Pose [°]	**2.29±1.31**	10.35±7.8	4.36±0.82	5.44±3.4
Age [yo]	**2.02±1.38**	14.63±8.4	14.38±8.5	7.11±6.1
Exp.	3.68±0.7	4.41±0.8	4.36±0.8	**2.94±0.6**
Illum.	**0.32±0.13**	0.62±0.21	0.61±0.21	**0.32±0.14**
Hair c.	**0.13±0.18**	0.33±0.24	0.24±0.18	0.15±0.14
ID preservation				
Same↓	**0.68±0.19**	0.82±0.3	0.97±0.35	1.16±0.34
Not same↑	**1.9±0.24**	1.78±0.23	1.79±0.25	1.7±0.26
FID ↓				
FID	5.72	6.48	9.1	**3.32**

Table 5: **Ablation study:** Comparison of our method vs. training end-to-end (single phase) and vs. using a non-disentangled StyleGAN2.

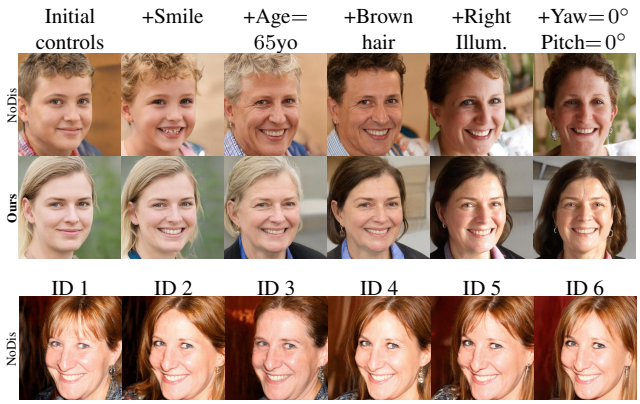

Figure 8: **Ours vs. NoDis:** In row 1 (NoDis) and row 2 (ours), from left to right, each column changes one control. The ID input is not changed. In row 3 (NoDis), each column has a different ID input and same control inputs.

FID metrics. As expected, the E2E-10x model achieves better control precision than the E2E model at the expense of reduced photorealism (FID score) and ID preservation. Nonetheless, at both ends of the spectrum the results are inferior to those achieved by our two-phased model. We present qualitative comparisons in the supplementary. Table 5 indicates that NoDis does not preserve ID. This is backed up by the first row of Fig. 8. In the third row of Fig. 8 we show images generated for different ID vectors but with the same set of controls. The mild variation in perceived ID demonstrates that the entanglement limits the possible IDs, given a set of controls. Moreover, for NoDis the control precision is inferior except for the expression. We hypothesize that in order to reach a desired control, the model partially "adjusts the ID". This is most prominent for expression where the geometry of the face changes. Thus with limited ID preservation, it is "easier" to achieve a desired expression.

4.4. Disentangled projection of real images

We leverage the explicit control of our model for real image editing. To this end, we use latent space optimization to find a latent vector that corresponds to an input image. By naïvely following the projection method described in StyleGAN2 (Appendix D), the reconstructed image visually looks different from the input image. A remedy to this phenomenon proposed in [5] is to project the image to an extended latent space, w^+, such that each resolution level has its own latent vector. We find this approach is indeed useful for accurate reconstruction. However, when we modified the different sub-vectors, we observed a strong deterioration in the image quality and a change in other unmodified attributes. In absence of explicit constraints on the feasible solutions' space, two different issues arise: (1) part of the sub-vectors end-up encoding a semantically different information from the one they were intended for, *e.g.*, the pose latent vector may encode some information of the ID or the expression, and (2) the reconstructed latent vector may not lie in the semantically meaningful manifold. A similar phenomenon was reported in Zhu *et al*. [56]. As a mitigation to the above, we introduce two changes. First, rather than extending the entire $\mathcal{W}$ space, we only extend the $\mathcal{W}^{ID}$ and $\mathcal{W}^{other}$ sub-spaces. Second, we constrain the remaining sub-vectors to reside on approximated linear subspaces of their corresponding manifolds. We achieve this using the following approach: we perform PCA for each latent subspace of 10K randomly sampled sub-vectors $\mathbf{w}$, where the number of components are selected so as to preserve 50% of the variance. During the optimization process, we project the latent sub-vectors to the truncated PCA spaces and reproject them back to the corresponding spaces. Once we find the corresponding latent vector, we can edit the image by modifying attribute k latent sub-vector, using E_k. We provide an ablation study of the proposed changes in the supplementary material.

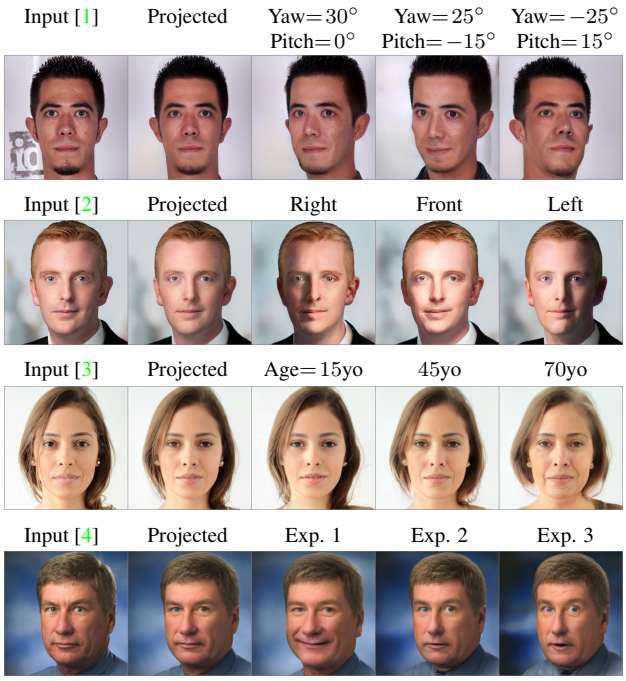

Figure 9: **Disentangled Projection:** The two leftmost columns refer to the input and projected images, respectively. The remaining columns demonstrate editing results of pose, illumination, age and expression.

In Fig. 9 we show real images, their projections and the result of editing their attributes. While the projected image does not achieve a perfect reconstruction, the disentanglment of the latent space is preserved, allowing for an explicit control of the desired attributes without affecting others. In the second row of Fig. 9 we can see that the GAN can accurately model the shadows on the face's curvature and skin folds as well as model the reflection of the light source in the person's eyes. This implies the GAN learns a latent 3D representation of the faces.

5. Conclusions

We proposed a novel framework for training GANs in a disentangled manner, that allows explicit control over generation attributes. For a variety of attributes, a predictor of that attribute is enough to achieve explicit control over it. Our method extends the applicability of explicitly controllable GANs to additional domains other than human faces. The GAN is complemented by a real image projection method that projects images to a disentangled latent space, maintaining explicit control. We believe this work opens up a path for improving the ability to control general-purpose GAN generation. Additional details can be found at alonshoshan10.github.io/gan_control/.

References

[1] The original image is at http://www.flickr.com/photos/quakecon/3923570806 and is licensed under: http://www.creativecommons.org/licenses/by/2.0. 8

[2] The original image is at http://www.flickr.com/photos/dereknolan/5309847731 and is licensed under: http://www.creativecommons.org/licenses/by/2.0. 8

[3] The original image is at http://www.flickr.com/photos/67548743@N02/6854926480 and is licensed under: http://www.creativecommons.org/licenses/by/2.0. 8

[4] The original image is at http://www.flickr.com/photos/ugacommunications/6005899336 and is licensed under: http://www.creativecommons.org/licenses/by-nc/2.0. 8

[5] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to Embed Images Into the StyleGAN Latent Space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 8

[6] Rameen Abdal, Peihao Zhu, Niloy Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows, 2020. 3

[7] Guha Balakrishnan, Yuanjun Xiong, Wei Xia, and Pietro Perona. Towards Causal Benchmarking of Bias in Face Analysis Algorithms. 2020. 2

[8] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Towards open-set identity preserving face synthesis. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6713–6722. IEEE Computer Society, 2018. 3

[9] Volker Blanz and Thomas Vetter. A Morphable Model for the Synthesis of 3D Faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, pages 187–194, 1999. 1, 2, 4

[10] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations*, 2019. 2, 5

[11] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR*, 2018. 3

[12] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: Diverse Image Synthesis for Multiple Domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR*, 2020. 3

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 7

[14] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 4, 6

[15] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and Controllable Face Image Generation via 3D Imitative-Contrastive Learning. In *IEEE Computer Vision and Pattern Recognition*, 2020. 2, 3, 5

[16] 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 Workshops*, 2019. 4

[17] Xin Ding, Yongwei Wang, Zuheng Xu, William J Welch, and Z. Jane Wang. Cc{gan}: Continuous conditional generative adversarial networks for image generation. In *International Conference on Learning Representations*, 2021. 2

[18] Chris Donahue, Zachary C Lipton, Akshay Balsubramani, and Julian McAuley. Semantically Decomposing the Latent Spaces of Generative Adversarial Networks. In *ICLR*, 2018. 2

[19] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3D Morphable Face Models—Past, Present, and Future. *ACM Transactions on Graphics (TOG)*, 39(5):1–38, 2020. 1

[20] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image Style Transfer Using Convolutional Neural Networks. In *CVPR*, 2016. 7

[21] 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, 2

[22] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering Interpretable GAN Controls. *arXiv preprint arXiv:2004.02546*, 2020. 2

[23] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *IEEE Trans. Image Process.*, 28(11):5464–5478, 2019. 3

[24] 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 *NIPS*, 2017. 5

[25] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal Unsupervised Image-to-Image Translation. In *ECCV*, 2018. 3

[26] Ali Jahanian, Lucy Chai, and Phillip Isola. On the "Steerability" of Generative Adversarial Networks. In *International Conference on Learning Representations*, 2020. 2

[27] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive Growing of GANs for Iimproved Quality, Stability, and Variation. *arXiv preprint arXiv:1710.10196*, 2017. 1, 2

[28] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *arXiv preprint arXiv:2006.06676*, 2020. 6

[29] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR*, pages 4401–4410, 2019. 1, 2, 4, 5

[30] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and Improving the Image Quality of StyleGAN. *CoRR*, abs/1912.04958, 2019. 1, 2, 3, 5

[31] Marek Kowalski, Stephan J. Garbin, Virginia Estellers, Tadas Baltrušaitis, Matthew Johnson, and Jamie Shotton. CONFIG: Controllable Neural Face Image Generation. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 5, 6

[32] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a Model of Facial Shape and Expression from 4D Scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 1

[33] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, 2014. 2

[34] Takeru Miyato and Masanori Koyama. cgans with projection discriminator. *ArXiv*, abs/1802.05637, 2018. 2

[35] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. HoloGAN: Unsupervised Learning of 3D Representations From Natural Images. In *The IEEE International Conference on Computer Vision (ICCV)*, Nov 2019. 5

[36] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *ICML*, 2017. 2

[37] Ayellet Tal Ori Nizan. Breaking the Cycle - Colleagues are All You Need. In *IEEE Conference on Computer Vision and Pattern Recognition CVPR*, 2020. 3

[38] Xingang Pan, Bo Dai, Ziwei Liu, Chen Change Loy, and Ping Luo. Do 2d {gan}s know 3d shape? unsupervised 3d shape reconstruction from 2d image {gan}s. In *International Conference on Learning Representations*, 2021. 3

[39] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic Image Synthesis with Spatially-Adaptive Normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR*, 2019. 3

[40] R. Ramamoorthi and P. Hanrahan. An Efficient Representation for Irradiance Environment Maps. In *SIGGRAPH '01*, 2001. 4

[41] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep Expectation of Apparent Age from a Single Image. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, December 2015. 4

[42] Nataniel Ruiz, Eunji Chong, and James M. Rehg. Fine-Grained Head Pose Estimation Without Keypoints. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018. 4

[43] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. Learning to Regress 3D Face Shape and Expression from an Image without 3D Supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7763–7772, 2019. 1

[44] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the Latent Space of GANs for Semantic Face Editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[45] 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, CVPR 2018*, 2018. 2

[46] Yujun Shen, Bolei Zhou, Ping Luo, and Xiaoou Tang. Facefeat-gan: a two-stage approach for identity-preserving face synthesis. *CoRR*, abs/1812.01288, 2018. 3

[47] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *3rd International Conference on Learning Representations, ICLR*, 2015. 7

[48] Henrique Siqueira, Sven Magg, and Stefan Wermter. Efficient Facial Feature Learning with Wide Ensemble-based Convolutional Neural Networks, Feb 2020. 4

[49] Nurit Spingarn, Ron Banner, and Tomer Michaeli. {GAN} "steerability" without optimization. In *International Conference on Learning Representations*, 2021. 3

[50] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Perez, Michael Zollhofer, and Christian Theobalt. StyleRig: Rigging StyleGAN for 3D Control Over Portrait Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[51] Ayush Tewari, Mohamed Elgharib, Mallikarjun BR, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zöllhofer, and Christian Theobalt. Pie: Portrait image embedding for semantic control. volume 39, December 2020. 3

[52] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR*, 2018. 3

[53] Ceyuan Yang, Yujun Shen, and Bolei Zhou. Semantic Hierarchy Emerges in Deep Generative Representations for Scene Synthesis. *arXiv preprint arXiv:1911.09267*, 2019. 2

[54] Xu Yao, Gilles Puy, Alasdair Newson, Yann Gousseau, and Pierre Hellier. High resolution face age editing. *arXiv preprint arXiv:2005.04410*, 2020. 3

[55] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid Scene Parsing Network. In *CVPR*, 2017. 4

[56] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain GAN Inversion for Real Image Editing. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020. 3, 8

[57] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *European conference on computer vision*, pages 597–613. Springer, 2016. 3

[58] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *IEEE International Conference on Computer Vision ICCV*, 2017. 3

[59] Peiye Zhuang, Oluwasanmi O Koyejo, and Alex Schwing. Enjoy your editing: Controllable {gan}s for image editing via latent space navigation. In *International Conference on Learning Representations*, 2021. 3