Multi-Directional Subspace Editing in Style-Space

Chen Naveh
School of Computer Science, Reichman University
navehchen1@gmail.com

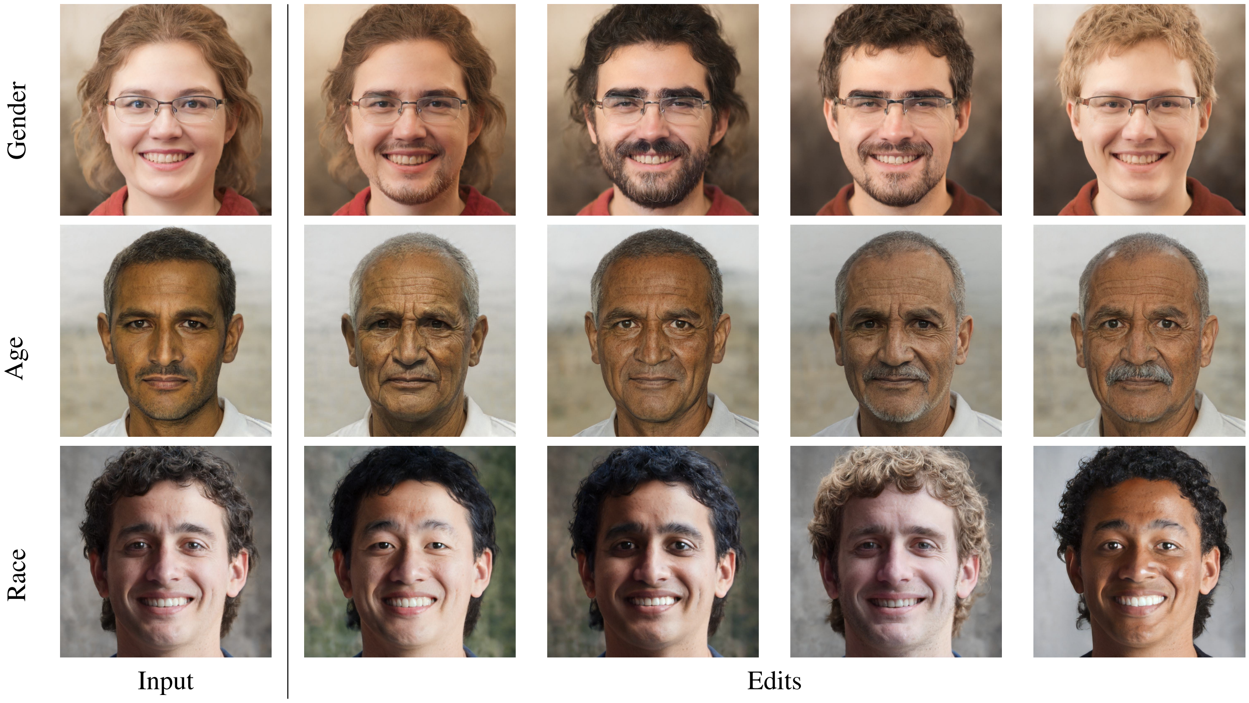

Figure 1: Multi-directional human face editing. Each row indicates changes in a subspace associated with a particular attribute: (top to bottom) gender, age and race.

Abstract

This paper describes a new technique for finding disentangled semantic directions in the latent space of StyleGAN. Our method identifies meaningful orthogonal subspaces that allow editing of one human face attribute, while minimizing undesired changes in other attributes. Our model is capable of editing a single attribute in multiple directions, resulting in a range of possible generated images. We compare our scheme with three state-of-the-art models and show that our method outperforms them in terms of face editing and disentanglement capabilities. Additionally, we suggest quantitative measures for evaluating attribute separation and disentanglement, and exhibit the superiority of our model with respect to those measures[1].

1. Introduction

Recent developments in computer vision have enabled the generation of photorealistic, high-resolution synthetic images. The most notable technique is Generative Adversarial Networks (GANs) [11], which have advanced the progress of many applications, including image generation, super-resolution, image inpainting, and more. In particular, the generator of StyleGAN [1], one of the most notable GAN models, has been extensively explored by researchers.

StyleGAN samples a latent vector $\mathbf{z} \in \mathbb{R}^{512}$ from a

[1] Project page and code are available at https://chennaveh.github.io/MDSE/

Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and maps it to an intermediate vector $\mathbf{w} \in \mathcal{W} = \mathbb{R}^{512}$ using a mapping network. This vector is then used to generate a 1024x1024 RGB image. The vector $\mathbf{w}$ is inserted into 18 multi-resolution style blocks that control various characteristics of the synthesized image. Vectors at lower resolutions determine high-level features such as pose and hair. Intermediate resolutions affect facial expressions, while vectors at higher resolutions dictate fine details like colors and texture. The original StyleGAN model uses the $\mathcal{W}$ latent space and the same $\mathbf{w}$ vector for all 18 style blocks. Subsequent studies [1, 2, 26, 32] utilize the $\mathcal{W}^+ = \mathbb{R}^{18\cdot512}$ space, which extends the $\mathcal{W}$ space, and applies different $\mathbf{w}$ vectors to different style blocks. This allows for enhanced control over the resulting image. The $\mathcal{W}^+$ latent space is commonly used for inverse mapping, converting an image into $\mathbf{w} \in \mathcal{W}^+$.

Recent works have explored methods for gaining control over the synthesized images by altering the latent space vectors [3, 7, 12, 25, 28, 29]. For instance, shifting the latent vector of a generated image towards the direction corresponding to "smile" can progressively amplify the smile in the output image. While these methods showcase impressive and realistic editing capabilities, pinpointing the attribute directions remains a challenge. One approach is to use vectors in latent space sampled from the GAN model combined with attribute ground truth (either manually labeled or determined using a pre-trained attribute estimator) [10, 16, 28]. A linear classifier in the latent space is then employed. The normal to the classification boundary corresponds to the most significant direction of the attribute. These models often suffer from issues of entangled data and biases in the training set. Such biases might result in attribute correlations, like glasses with age or beard with gender. Consequently, adding glasses to a face may inadvertently age it. Other methods use unsupervised techniques [12, 29], such as PCA, to find meaningful orthogonal directions in the latent space. However, these directions typically require subjective human post-annotation to link directions to attributes in the synthesized images. All these models share the notion of a singular direction for each attribute, despite the possibility of some attributes defined by multiple dimensions (*e.g*., age might be affected by factors like hair color and skin wrinkles, see Fig. 1).

In this work, we present MDSE, an acronym for *Multi-Directional Subspace Editing*. This framework aims to identify orthogonal semantics within the latent space of a pre-trained StyleGAN. Our goal is to discover meaningful subspaces that are mutually orthogonal, with each subspace controlling specific facial attributes. "Multi-directional editing" means altering a latent vector within a particular subspace in various directions. This approach enables the generation of a wide range of images, each varying in a specific attribute. Furthermore, because the subspaces are mutually orthogonal, modifications in one attribute result in minimal changes to others. We explore the disentanglement capabilities of our model and quantitatively assess its performance in comparison to leading image editing techniques.

2. Related Work

2.1. Generative Adversarial Networks (GANs)

Generative models have revolutionized the generation of vast amounts of visual content. They can produce visually compelling content with remarkable fidelity, making them indispensable for image editing applications. GAN models learn a mapping from a specific distribution (usually Gaussian or uniform) to the target distribution. The generator $G(\cdot)$ is trained to fool an adversary $D(\cdot)$, which aims to differentiate between the generated images and the real ones. Both models $G(\cdot)$ and $D(\cdot)$ are trained concurrently using the min-max loss [11]. The GAN approach is widely recognized and often outperforms other techniques, such as variational autoencoders (VAEs) [21]. Recently, denoising diffusion probabilistic models (DDPM) [15] have risen in prominence and showcased superior results. However, despite their leading performance in image generation, their potential in image editing and semantic interpretation remains relatively unexplored. In this work, our attention is centered on a GAN-based generator, specifically the StyleGAN [18–20], which exhibits exceptional capability in the human face domain.

GAN Latent Space. Given a pre-trained generator $\mathbf{x} = G(\mathbf{w})$, GAN inversion $\hat{\mathbf{w}} = G^{-1}(\mathbf{x})$ is the process of inverting a specified image $\mathbf{x}$ into a latent vector $\hat{\mathbf{w}}$ that faithfully reconstructs the image using the generator. In other words, $G(\hat{\mathbf{w}}) \approx \mathbf{x}$. Most of the common methods either train an encoder [24, 26, 32] from the image space into the latent space or optimize the latent vector, using backpropagation, directly until it generates the desired image [1, 2, 22]. Recent approaches that modify the generator's weight have gained traction, resulting in enhanced image reconstruction [5, 27]. Previous works show that the latent space of a pre-trained GAN model encodes image semantics in a meaningful structure. Hence, latent codes corresponding to images with similar semantics tend to cluster closely in the latent space [28]. For instance, within the domain of human face generation, faces sharing characteristics (*e.g*., being young and blond) will have nearby latent vectors. This foundational insight drives modern image editing techniques.

2.2. Multi-Directional Manipulation in Latent Space

The idea of editing an image through latent space manipulation has been extensively explored in recent years [7, 25]. Such approaches aim to harness the strong capabilities of the generator. Therefore, they keep the generator network static and conduct vector operations solely in the latent space. Recent works primarily rely on a linear latent space assumption and demonstrate pleasing-looking results of image editing by adding or subtracting vectors within the latent space [12, 28, 29]. InterFaceGAN [28] employs pre-trained binary classifiers to label images with facial features, such as young-old, male-female, and with/without glasses, among others. It then trains a linear classifier (specifically, a Support Vector Machine or SVM) on the respective latent vectors to derive a classifying hyperplane. The normal to this hyperplane serves as the 1D direction for editing the relevant attribute.

GANSpace [12] adopts a data-driven approach, applying PCA on a set of vectors sampled from $\mathcal{W}$ to find meaningful directions. The eigenvectors associated with the largest eigenvalues serve as the editing directions. Some derived directions can be entangled, such as head rotation and gender. To address this issue, edits are restricted to a specifically selected subset of the generator's layers.

Another work, SeFa [29], avoids the image sampling pre-processing step. Instead, it determines a closed-form factorization of the latent space by utilizing an eigen-decomposition of the generator's weights. For StyleGAN, the style affine transformation matrices are employed for vector decomposition. The eigenvectors corresponding to the greatest variations are selected to define the semantic directions. However, this unsupervised method faces two main issues. Firstly, there's a subjective post-annotation process to match the directions with relevant semantics. Secondly, the directions that are derived often mix several semantics (*e.g*., age and glasses), a result of biases present in the training data.

A distinct set of studies moves away from the linear approach, opting instead to learn non-linear functions to manipulate latent vectors. StyleFlow [3] employs continuous normalizing flows conditioned on the image attribute vector. This attribute vector must be provided using pre-trained networks before every edit. Another non-linear approach is StyleRig [30], which incorporates a 3D semantic network and harnesses it to perform rigid edits such as pose and light. Although non-linear approaches generally perform better, our research primarily emphasizes the more comprehensive linear approach.

All previous methods that manipulate the latent space of StyleGAN define a semantic as a singular direction. Both linear and non-linear approaches ultimately derive positive and negative directions for editing a desired attribute. We believe that this oversimplification may restrict the editing capabilities and limit the diversity of the generated images. For example, imagine we want to edit the gender of a given face. One person may argue that hairstyle is a distinguishing factor between men and women, while another might think of facial structure or the presence of facial hair as a more prominent factor. In this work, we overcome this variability by associating an attribute with an n-dim **subspace** where $n \geq 1$, containing more than a single direction, in the semantic space. We call this a multi-directional subspace, as different vectors within the subspace can affect the resulting image differently, but all affect the associated attribute (see examples in Fig. 1). We also require different attributes to be associated with mutually orthogonal subspaces, to promote disentanglement between attributes.

We outline our contributions as follows:

- We propose MDSE to extend the concept of meaningful latent directions to multi-directional subspaces, thereby diversifying and expanding the capabilities of the editing process.
- MDSE identifies orthogonal directions in StyleGAN's latent space, to promote disentanglement. We introduce an orthogonality loss in the model training and demonstrate its significance in preserving some image attributes while editing others. We visualize the results and demonstrate improved results in consecutive edits of multiple attributes (refer to Fig. 4).
- We develop a new metric for evaluating the disentanglement properties of image editing. In addition to the conventional approaches of visual comparisons, we suggest using our evaluation metrics as a quantitative measurement of disentanglement capabilities.

3. Method

Given a set of attributes, $B \triangleq \{gender, glasses, \cdots\}$, we aim to find a subspace for each attribute $b_i \in B$ so that editing within this subspace leads to changes in the associated attribute b_i. Different directions in each subspace are responsible for distinct visual semantics. To achieve disentanglement, we desire that each subspace solely influences its respective attribute.

Certain attributes, such as glasses and age, have been observed to correlate with each other [28], making disentangled editing challenging. To overcome this problem, our model seeks to decompose the latent space $\mathcal{W}^+$ into multiple orthogonal subspaces each of which is associated with a single attribute. This requirement has two main outcomes. Firstly, editing within a subspace facilitates multi-directional edits of a single attribute. Secondly, the orthogonality ensures that altering a specific facial attribute doesn't affect other attributes.

3.1. Latent Space Decomposition

We decompose $\mathcal{W}^+$ into $N+1$ orthogonal subspaces $\{S_i\}_{i=0}^{N}$ and define each subspace S_i as

$$S_i \triangleq span\left\{\mathbf{p}_i^1, \mathbf{p}_i^2, \cdots, \mathbf{p}_i^{n_i}\right\}, n_i < dim(\mathcal{W}^+) \quad (1)$$

where $\mathbf{p}_i^j \in \mathcal{W}^+$, and $\{\mathbf{p}_i^j\}_{j=1}^{n_i}$ are linearly independent vectors. Therefore, $\{\mathbf{p}_i^j\}_{j=1}^{n_i}$ form a basis for subspace S_i with cardinality n_i. There are $N+1$ subspaces, where $N = |B|$ is the size of the attribute set B. Each S_i corresponds to b_i respectively, while S_0 is associated with all other information that is not labeled in B. This may include other semantics, *e.g.*, clothing and image background. The correspondence of S_i and b_i will be elaborated on later in the paper.

To ensure orthogonality, we require the following conditions:

$$S_i \subset \mathcal{W}^+ \ , \forall i \in \{0, \cdots, N\} \tag{2}$$

$$\bigoplus_{i=0}^{N} S_i = \mathcal{W}^+ \tag{3}$$

$$S_i \perp S_j \ , \forall i \neq j \tag{4}$$

where $\bigoplus$ denotes direct sum, and $\perp$ signifies orthogonal subspaces. Notice that the entire vector set $\{\mathbf{p}_i^j\}$ forms a basis for $\mathcal{W}^+$. As a result, we can uniquely express every vector $\mathbf{w} \in \mathbb{R}^{18 \cdot 512}$ with a set of scalars $\{a_i^j\}$ and represent $\mathbf{w}$ as the following linear combination:

$$\mathbf{w} = \sum_{i=0}^{N} \sum_{j=1}^{n_i} a_i^j \mathbf{p}_i^j \tag{5}$$

This expression can also be represented in matrix form:

$$\mathbf{w} = \left[\underbrace{\begin{bmatrix} | & & | \\ \mathbf{p}_0^1 & \cdots & \mathbf{p}_0^{n_0} \\ | & & | \end{bmatrix}}_{\mathbf{P}_0} \cdots \underbrace{\begin{bmatrix} | & & | \\ \mathbf{p}_N^1 & \cdots & \mathbf{p}_N^{n_N} \\ | & & | \end{bmatrix}}_{\mathbf{P}_N} \right] \begin{bmatrix} \begin{bmatrix} a_0^1 \\ \vdots \\ a_0^{n_0} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} a_N^1 \\ \vdots \\ a_N^{n_N} \end{bmatrix} \end{bmatrix} \tag{6}$$

$$\mathbf{w} = \sum_{i=0}^{N} \mathbf{P}_i \mathbf{a}_i = \underbrace{\left[\mathbf{P}_0, \cdots, \mathbf{P}_N\right]}_{\mathbf{P}} \underbrace{\begin{bmatrix} \mathbf{a}_0 \\ \vdots \\ \mathbf{a}_N \end{bmatrix}}_{\mathbf{a}} \tag{7}$$

where $\mathbf{P}$ is a matrix defined by:

$$\mathbf{P} = \left[\mathbf{p}_0^1 \cdots \mathbf{p}_0^{n_0} \cdots \mathbf{p}_N^1 \cdots \mathbf{p}_N^{n_N}\right] \tag{8}$$

and $\mathbf{a}^T = [\mathbf{a}_0^T, \ldots, \mathbf{a}_N^T]$ is a vector of coefficients. Finding $\mathbf{P}$ that satisfies Eq. (2)-(4) is the core part of our framework.

3.2. Training Procedure

Our dataset consists of a set of 2,000 vector pairs $\left\{(\mathbf{w}^{(i)}, \mathbf{y}^{(i)})\right\}_{i=1}^{2000}$. The latent vectors $\left\{\mathbf{w}^{(i)}\right\}$ are generated from random vectors sampled from a Gaussian distribution $\mathbf{z}^{(i)} \sim \mathcal{N}(0,1)$ and then mapped to $\mathcal{W}^+$ space using the StyleGAN mapping function: $\mathbf{w}^{(i)} = M(\mathbf{z}^{(i)})$. Each sample $\mathbf{w}^{(i)} \in \mathbb{R}^{18 \cdot 512}$ is mapped onto the image space using the StyleGAN generator, $\mathbf{x}^{(i)} = G(\mathbf{w}^{(i)})$ and then annotated using pre-trained classifiers to determine an attribute score vector:

$$\mathbf{y}^{(i)} \triangleq (y_1^{(i)}, \cdots, y_N^{(i)}) = \Big(\mathcal{C}_1(\mathbf{x}^{(i)}), \cdots, \mathcal{C}_N(\mathbf{x}^{(i)})\Big) \tag{9}$$

Here, each $y_k^{(i)}$ denotes the b_k attribute score for sample i. Depending on the attribute, $y_k^{(i)}$ is either a discrete or continuous number. $\mathcal{C}_k$ represents the pre-trained classifier for attribute k. For age, smile, gender, and glasses, we used the face attribute classifier from [13] trained on the FFHQ dataset [19]. For the pose attribute, we utilized img2pose [6] for face estimation. Additionally, we employed a race classifier [17] trained on the Yahoo YFCC100M dataset [31].

The primary goal during the training phase is to determine the matrix $\mathbf{P}$ and use it to reconstruct all samples from the training set. To satisfy Eq. (5), we jointly learn a vector $\mathbf{a}^{(i)}$ for each vector $\mathbf{w}^{(i)}$ such that $\mathbf{w}^{(i)} = \mathbf{P}\mathbf{a}^{(i)}$. We then introduce the following loss:

$$\mathcal{L}_{rec}^{(i)} = \left\| \mathbf{w}^{(i)} - \mathbf{P}\mathbf{a}^{(i)} \right\|_1 \tag{10}$$

where $\mathcal{L}_{rec}$ represents the reconstruction loss. Instead of pixel-based image reconstructions, we use a L_1 loss in the latent space using the original vector $\mathbf{w}^{(i)} = M(\mathbf{z}^{(i)})$ as the target value.

Additionally, to enforce disentanglement, as referred in Eq. (4), we introduce an orthogonality loss:

$$\mathcal{L}_{orth} = \sum_{i \neq j} \left\| \mathbf{P}_i^T \mathbf{P}_j \right\|_2^2 \tag{11}$$

and require that the columns of matrices $\mathbf{P}_i, \mathbf{P}_j$ be orthogonal for $i \neq j$, where $\|\cdot\|_2$ is the element-wise Frobenius-norm.

Since learning disentangled representations is fundamentally impossible without a supervised inductive bias on the data [23], we leverage the attribute vector $\mathbf{y}^{(i)}$ and introduce a mixing loss to establish the association between S_k and b_k. Given a vector $\mathbf{w}^{(i)}$, we can import attribute b_k from a randomly chosen vector $\mathbf{w}^{(j)}$, to generate $\mathbf{w}_{mix}^{(i)}$:

$$\mathbf{w}_{mix}^{(i)} = \mathbf{P}_k \mathbf{a}_k^{(j)} + \sum_{l \neq k} \mathbf{P}_l \mathbf{a}_l^{(i)} \tag{12}$$

and its corresponding image $\mathbf{x}^{(i)}_{mix} \triangleq G(\mathbf{w}^{(i)}_{mix})$. For the modified image we require $\mathcal{C}_k(\mathbf{x}^{(i)}_{mix})$ to be similar to $y^{(j)}_k$ while keeping the other attributes unchanged. Thus a mixing loss is added to the network to force changes only in b_k:

$$\mathcal{L}^{(i)}_{mixing} = L_k\left(\mathcal{C}_k(\mathbf{x}^{(i)}_{mix}), y^{(j)}_k\right) + \sum_{l\neq k} L_l\left(\mathcal{C}_l(\mathbf{x}^{(i)}_{mix}), y^{(i)}_l\right) \tag{13}$$

L_i are loss functions that vary based on the attribute types. For categorical labels, we use softmax with cross-entropy loss, whereas continuous labels are optimized using the L_1 loss. In practice, we mix all attributes together from randomly chosen vectors. This introduces complex changes to the image and encourages the association of a single subspace with a single attribute.

Finally, our model is trained using an objective function comprised of the three losses:

$$\mathcal{L} = \lambda_{orth}\mathcal{L}_{orth} + \frac{1}{n}\sum_i \mathcal{L}^{(i)}_{rec} + \lambda_{mixing}\mathcal{L}^{(i)}_{mixing} \tag{14}$$

where λ_{orth}, λ_{mixing} are hyperparameters.

4. Experiments

In this section, we evaluate the performance of our model compared to state-of-the-art image editing models, all of which utilize the StyleGAN generator for image synthesis.

To test our model, we generated a source and target latent vector, projected them onto a subspace S_k, and replaced the projections of the source vector with those from the target. Since our subspaces are orthogonal, we expect to see changes only in the corresponding attribute b_k. Example results are presented in Fig. 2. Notably, if the source face wears glasses, these remain post-edit since we only incorporate pose, smile, and race from the target face.

4.1. Comparison with Previous Methods

We compared our model with three previous image editing methods: InterFaceGAN [28], StyleFlow [3], and SeFa [29]. We evaluated the editing and disentanglement qualities of these models for a common set of attributes supported by all of them, such as age, gender, smile, pose, and glasses. Additionally, since our method discovers a subspace rather than a single direction, to ensure a fair comparison, we trained a linear SVM inside each subspace. We then used the normal to the hyperplane as the direction for editing. This model is used in Section 4.2 and Section 4.3.

4.2. Qualitative Comparison

Figure 3 offers a comparative assessment of the quality of real image editing. We inverted the images into StyleGAN's latent space using HyperStyle [4]. Our observations indicate that multiple edits on a single image can significantly diminish its quality, making it more prone to visual artifacts. Moreover, due to biases in the FFHQ dataset [19], some attributes are more correlated in StyleGAN's latent space than others (*e.g.*, younger individuals are less likely to wear glasses). As depicted in Fig. 3, while most models can effectively modify a single attribute, our model outperforms others in terms of accuracy when changing all attributes simultaneously and better maintains each individual attribute edit. Sequential editing is illustrated in Fig. 4, showing that as edits progress, the images retain previous attributes without a loss in quality.

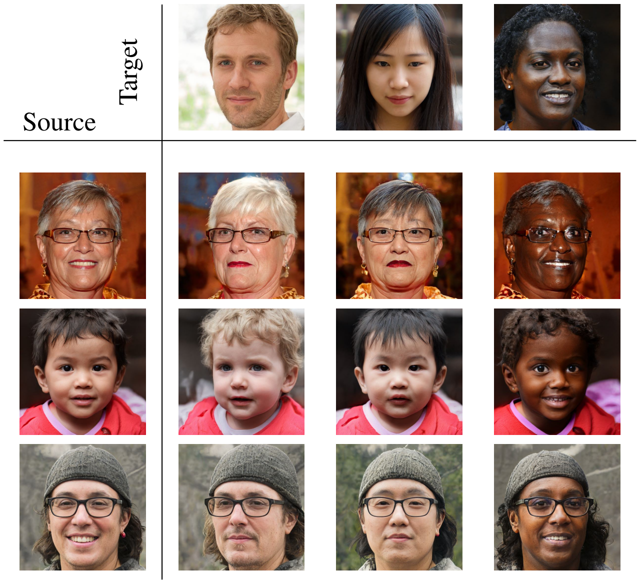

Figure 2: Image editing capabilities using source images and target attribute images. The attributes derived from the target images include pose, smile, and race.

In Fig. 1 we visualize our model's capability of generating diverse results for different attributes due to its multi-directional nature. The images are generated by shifting the latent vectors to different directions inside the relevant subspace. We found that some attributes (*e.g.*, smile, pose) behave like binary attributes, meaning they contain most of the information in a single direction. Others (*e.g.*, gender, age), however, can be edited in multiple directions resulting in various images. Additional results can be found in the supplementary material.

4.3. Quantitative Comparison

To evaluate our model's disentanglement capabilities, we develop and utilize two different methods: attribute correlation and face preservation. Furthermore, to measure the diversity introduced by multi-directional subspace editing, we utilize the LPIPS score [33] and the Frechet Inception Distance (FID) [14] to evaluate image fidelity.

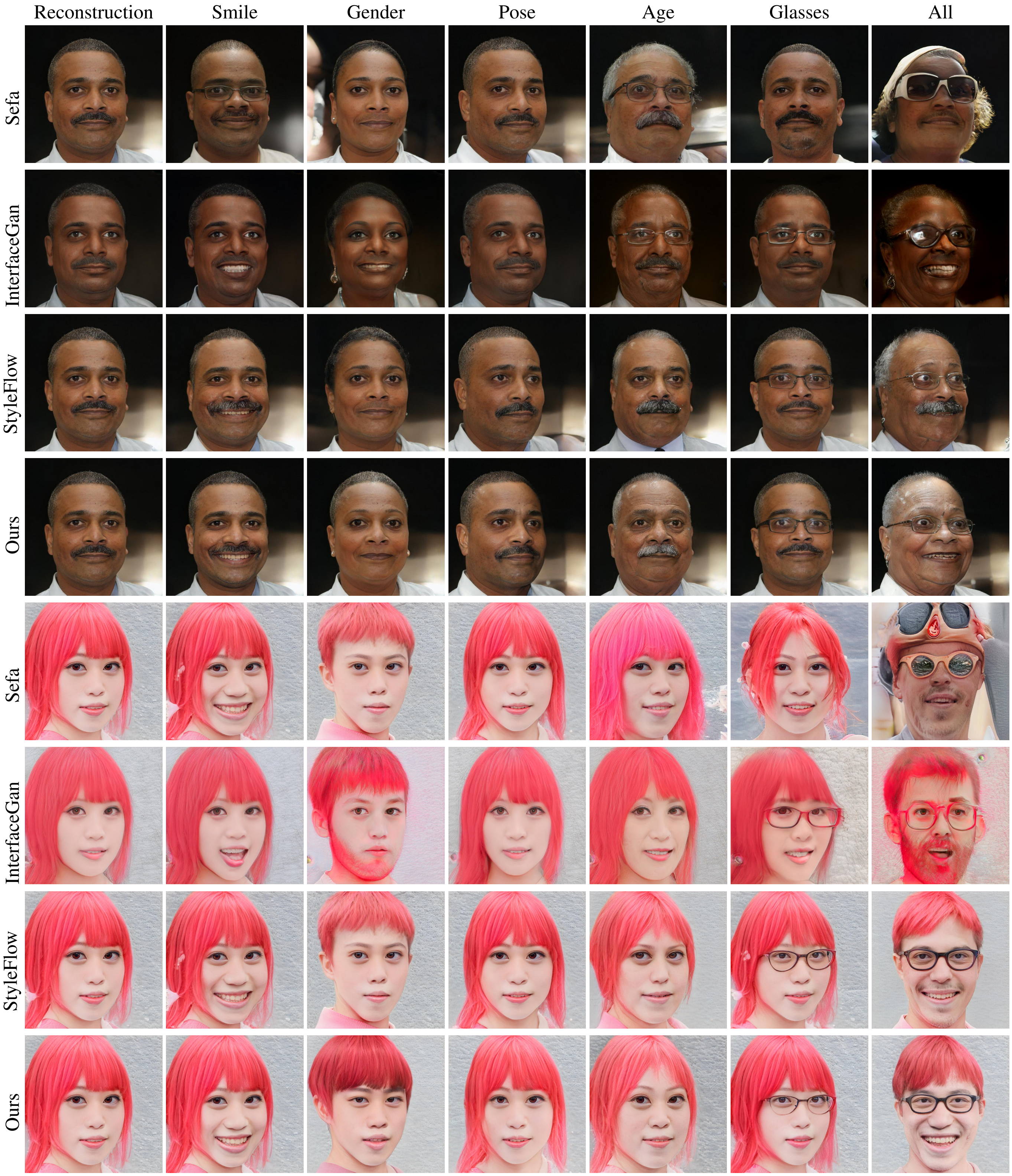

Figure 3: Comparison of real image editing between our method and the baseline approaches. The edit direction for each attribute was determined using a linear SVM classifier.

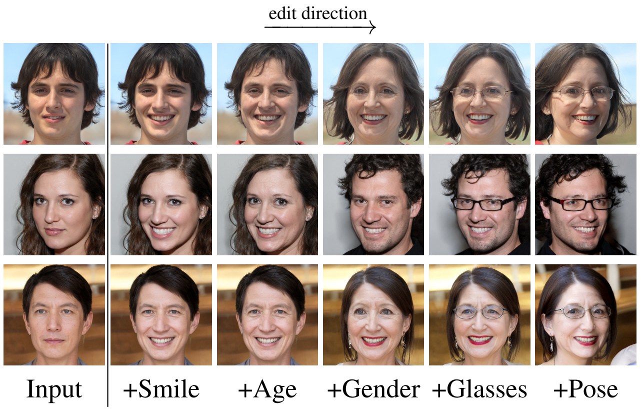

Figure 4: Sequential human face editing result.

Correlation between edited attributes: We can assign an attribute score to an image $\mathbf{x}$ using $\mathcal{C}_k$ as a feature extractor. The activation vector $l(\mathbf{x})$, derived from the last hidden layer of $\mathcal{C}_k$, is used to compute the distance between the feature vector $l(\mathbf{x})$ and the classifying hyperplane. This hyperplane is defined by $\left\{\mathbf{v} \mid \mathbf{c}_k^T \mathbf{v} + \mathbf{t}_k = 0\right\}$.

The attribute score of an image $\mathbf{x}$ is given by the distance of $l(\mathbf{x})$ from the separating hyperplane: $dist_k(\mathbf{x}) = \frac{\mathbf{c}_k^T l(\mathbf{x})+\mathbf{t}_k}{|\mathbf{c}_k|}$.

To assess the disentanglement capability of our model, we began by generating an evaluation set comprising $1K$ pairs of images paired with their desired face attributes. These images were directly sampled from the StyleGAN generator, while the attributes were independently generated for each image. For each image, we manipulated its latent vector $\mathbf{w} \in \mathcal{W}^+$ along the learned directions and fed it to the generator to produce an edited version, denoted as $\mathbf{x}_{edit}$. It's worth noting that in practice we change all attributes simultaneously.

Next, we measured the difference in attribute score between the original and the edited images. Taking the attribute "smile" as an example, we defined the perceptual distance as: $\Delta_{smile} = dist_{smile}(\mathbf{x}_{edit}) - dist_{smile}(\mathbf{x})$.

For the final assessment, we computed the Pearson correlation between the perceptual distances of every pair of attributes. Each entry in Table 1 displays the absolute correlation value. For instance, for the attributes "smile" and "pose", we determine the perceptual distances Δ_{smile} and Δ_{pose} across the entire evaluation set. The correlation value shown in the table is:

$$corr(pose, smile) = \left| \frac{cov(\Delta_{pose}, \Delta_{smile})}{E[\Delta_{pose}]E[\Delta_{smile}]} \right| \qquad (15)$$

Table 1 presents the correlation results of various methods. As previously mentioned [8, 12, 28, 29], we observe that certain attributes exhibit stronger correlations than others. For instance, the glasses-age correlation is consistently higher than the smile-pose correlation across all models.

Table 1: Attribute correlation matrices of edited images (in absolute values).

(a) SeFa

	Pose	Smile	Age	Gender	Glasses
Pose	1.000	0.150	0.039	0.133	0.051
Smile	0.150	1.000	0.001	0.170	0.233
Age	0.039	0.001	1.000	0.533	0.367
Gender	0.133	0.170	0.533	1.000	0.339
Glasses	0.051	0.233	0.367	0.339	1.000
Avg	0.093	0.115	0.235	0.293	0.247

(b) InterFaceGan

	pose	Smile	Age	Gender	Glasses
Pose	1.000	0.098	0.090	0.017	0.079
Smile	0.098	1.000	0.154	0.198	0.084
Age	0.090	0.154	1.000	0.565	0.600
Gender	0.017	0.198	0.565	1.000	0.363
Glasses	0.079	0.084	0.600	0.363	1.000
Avg	0.071	0.133	0.352	0.285	0.281

(c) StyleFlow

	Pose	Smile	Age	Gender	Glasses
Pose	1.000	0.113	0.126	0.103	0.056
Smile	0.113	1.000	0.264	0.011	0.074
Age	0.126	0.264	1.000	0.581	0.494
Gender	0.103	0.011	0.581	1.000	0.338
Glasses	0.056	0.074	0.494	0.338	1.000
Avg	0.099	0.115	0.366	0.258	0.240

(d) Ours

	Pose	Smile	Age	Gender	Glasses
Pose	1.000	0.018	0.008	0.029	0.046
Smile	0.018	1.000	0.120	0.050	0.100
Age	0.008	0.120	1.000	0.388	0.363
Gender	0.029	0.050	0.388	1.000	0.203
Glasses	0.046	0.100	0.363	0.203	1.000
Avg	0.025	0.072	0.219	0.167	0.178

The table's final row sums the off-diagonal columns, representing the average correlation score. Our method's lower scores highlight its superior disentanglement capabilities compared to the other techniques.

In our previous assessment, all attributes were adjusted simultaneously. To test the stability of attributes when only one is modified, we conducted an experiment editing one attribute at a time and measured the perceptual distances using the classifiers' final layers. Fig. 5 displays results from various methods. The x-axis represents changes in the edited attribute, while the y-axis indicates unintentional changes in other attributes. As edits become more extensive, the inter-attribute effect is more pronounced. Among all models, our approach shows the least deviation, indicating superior attribute disentanglement.

Table 2: Identity preservation scores by different models.

Edit	Metric	SeFa	InterFaceGan	StyleFlow	Ours
Smile	$C_s \Uparrow$	0.970	0.978	0.989	**0.991**
	$E_d \Downarrow$	0.329	0.276	0.192	**0.174**
Pose	$C_s \Uparrow$	0.978	0.979	0.981	**0.982**
	$E_d \Downarrow$	0.283	0.279	0.2662	**0.253**
Glasses	$C_s \Uparrow$	0.966	0.978	0.984	**0.985**
	$E_d \Downarrow$	0.348	0.270	0.227	**0.215**
All	$C_s \Uparrow$	0.911	0.933	0.935	**0.948**
	$E_d \Downarrow$	0.622	0.537	0.534	**0.467**

Notation: C_s - Cosine Similarity; E_d - Euclidean Distance.

Face Identity Preservation: Following the approach of the face identity score [3], we assess the edited images using a pre-trained face embedding network [9]. We restricted our edits to attributes that should preserve identity, such as pose, glasses, and smile. After editing, we embedded the images into a latent space using the network. We then calculated the Euclidean distance, E_d, and the cosine similarity, C_s, between the original and edited images.

Table 2 displays the results for face identity preservation. Our method consistently surpasses other techniques across all types of edits. While StyleFlow shows comparable scores for individual attribute edits, our model excels in preserving identity when multiple edits are combined.

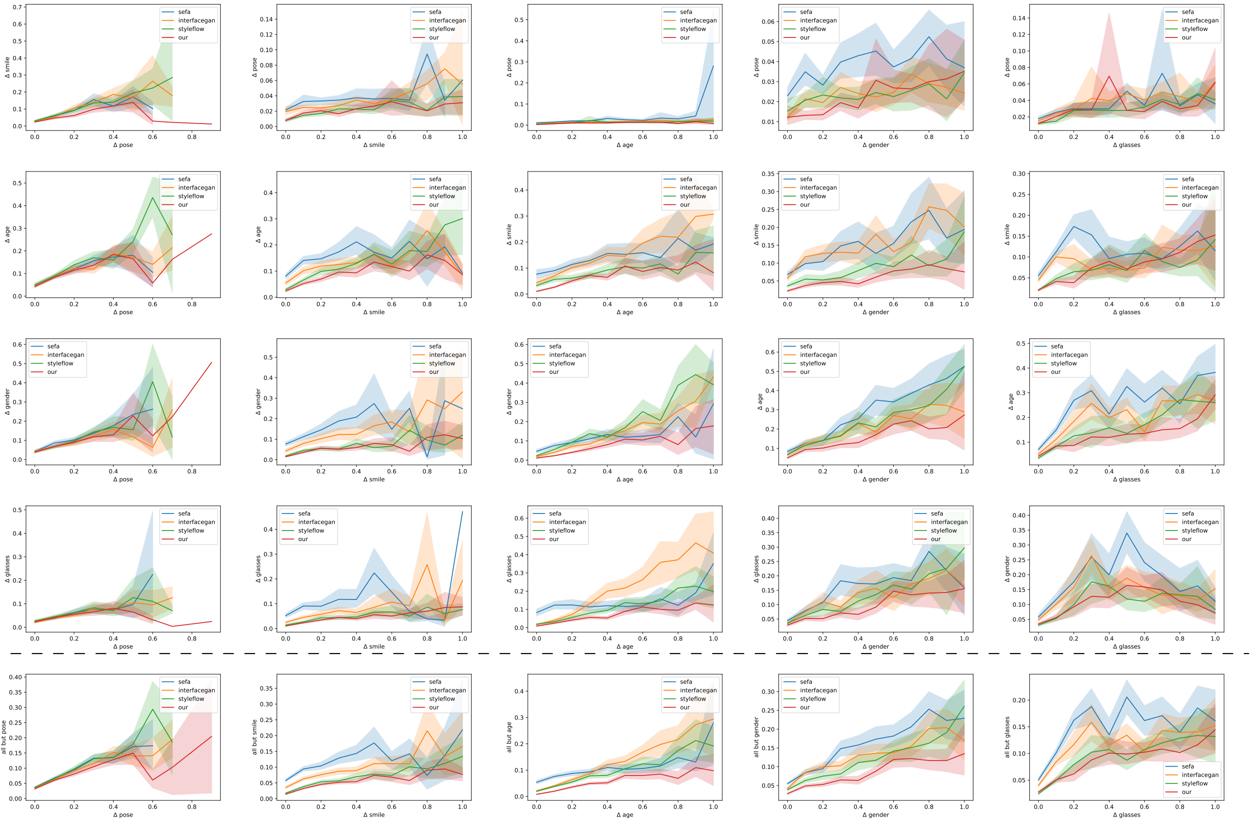

Figure 5: Comparative analysis of the inter-attribute effects exhibited by different models. The upper rows display how changes in one attribute vary as a function of changes in another attribute. The bottom row provides an average of all the rows above.

4.4. Diversity Analysis

In the previous sections, we introduced a single-direction variant of our model. Now, we will explore the performance of our model when multiple directions in the subspace are utilized for editing. To measure the diversity of the images produced by our model, we employed the LPIPS perceptual similarity score [33]. A lower LPIPS score indicates higher similarity between images, whereas a higher score signifies perceptual differences.

We generated a batch of 1,000 images and applied 5 random edits to each image. For every set of 5 edits, the similarity score between each pair of edited images was computed. Subsequently, all these scores were averaged to derive the final diversity score. We benchmarked our model against StyleFlow and the single-direction variant of our model, referred to as SVM, as detailed in Section 4.1. The FID score was also calculated to assess the visual quality of the generated images.

The results, summarized in Table 3, reveal that our approach — when multi-directional subspace editing is employed — not only fosters greater image diversity (as indicated by the LPIPS score) but also ensures superior visual quality (as denoted by the FID score). It's noteworthy that there is a marked enhancement in the multi-directional iteration of our model (subspace) when compared against its single-direction counterpart (SVM) or against StyleFlow.

Table 3: Diversity score results for edited images. Our multi-directional model achieves a higher score which indicates its ability to generate more diverse results while maintaining a lower FID score.

Method	LPIPS ⇑	FID ⇓
StyleFlow	0.213	27.467
Our (svm)	0.229	26.926
Our (subspace)	**0.321**	**23.648**

4.5. Ablation Study

To validate the importance of our orthogonality loss outlined in Section 3.2, we carried out an ablation study. We trained a version of our model with $\lambda_{orth} = 0$, keeping all other configurations, including losses and the training process, consistent. Post-training, we conducted consecutive edits starting from an original image, and the results can be viewed in Fig. 6.

We found that images edited with the orthogonality loss exhibit crisper details and fewer visual artifacts when subjected to multiple edits. We also carried out the same quantitative evaluations as detailed in Section 4.3. The results in Tables 4 and 5 indicate that our orthogonality loss leads to better attribute disentanglement in our model.

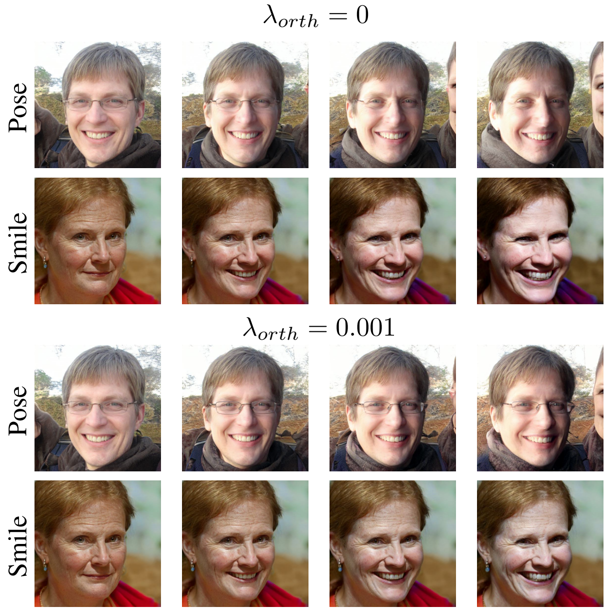

Figure 6: Comparison using orthogonality loss: First and third rows show reduced pose-glasses entanglement; second and fourth rows show better preservation of hair and shirt color.

Table 4: Attribute correlation matrices with and without orthogonal loss. The values represent the average correlation between an attribute and others.

Model	Pose	Smile	Age	Gender	Glasses
$\lambda_{orth} = 0$	0.149	0.242	0.342	0.283	0.253
$\lambda_{orth} = 0.001$	**0.025**	**0.072**	**0.219**	**0.167**	**0.178**

Table 5: Comparison of identity preservation using orthogonal loss.

Edit	Metric	$\lambda_{orth} = 0$	$\lambda_{orth} = 0.001$
All	C_s ⇑	0.932	**0.948**
	E_d ⇓	0.536	**0.467**

Notation: C_s - Cosine Similarity; E_d - Euclidean Distance.

5. Conclusions

In this work we proposed MDSE, a disentangling generative model for multi-attribute image editing. We introduce the concept of orthogonal subspaces that support multi-directional edits for diverse image generation. Additionally, this model effectively identifies disentangled latent subspaces, allowing precise control over the generated images and the displayed face attributes. We believe that integrating these concepts during the generator's training could further enhance the performance of disentangled models.

References

[1] 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*, pages 4432–4441, 2019. 1, 2

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8296–8305, 2020. 2

[3] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (TOG)*, 40(3):1–21, 2021. 2, 3, 5, 8

[4] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18511–18521, 2022. 5

[5] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit H Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. *arXiv preprint arXiv:2111.15666*, 2021. 2

[6] Vitor Albiero, Xingyu Chen, Xi Yin, Guan Pang, and Tal Hassner. img2pose: Face alignment and detection via 6dof, face pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7617–7627, 2021. 4

[7] Piotr Bojanowski, Armand Joulin, David Lopez-Paz, and Arthur Szlam. Optimizing the latent space of generative networks. *arXiv preprint arXiv:1707.05776*, 2017. 2

[8] Perla Doubinsky, Nicolas Audebert, Michel Crucianu, and Hervé Le Borgne. Multi-attribute balanced sampling for disentangled gan controls. *arXiv preprint arXiv:2111.00909*, 2021. 7

[9] A. Geitgey. face_recognition. https://github.com/ageitgey/face_recognition, 2021. 8

[10] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 5744–5753, 2019. 2

[11] 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

[12] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in Neural Information Processing Systems*, 33:9841–9850, 2020. 2, 3, 7

[13] Keke He, Yanwei Fu, Wuhao Zhang, Chengjie Wang, Yu-Gang Jiang, Feiyue Huang, and Xiangyang Xue. Harnessing synthesized abstraction images to improve facial attribute recognition. In *IJCAI*, pages 733–740, 2018. 4

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2

[16] Ali Jahanian, Lucy Chai, and Phillip Isola. On the" steerability" of generative adversarial networks. *arXiv preprint arXiv:1907.07171*, 2019. 2

[17] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021. 4

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

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2, 4, 5

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

[21] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[22] Zachary C Lipton and Subarna Tripathi. Precise recovery of latent vectors from generative adversarial networks. *arXiv preprint arXiv:1702.04782*, 2017. 2

[23] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019. 4

[24] Stanislav Pidhorskyi, Donald A Adjeroh, and Gianfranco Doretto. Adversarial latent autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14104–14113, 2020. 2

[25] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 2

[26] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2287–2296, 2021. 2

[27] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *arXiv preprint arXiv:2106.05744*, 2021. 2

[28] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representa-

tion learned by gans. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 2, 3, 5, 7

[29] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1532–1540, 2021. 2, 3, 5, 7

[30] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, 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*, pages 6142–6151, 2020. 3

[31] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 4

[32] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021. 2

[33] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5, 9