Pattern Recognition Letters
journal homepage: www.elsevier.com

Multi-Attribute Balanced Sampling for Disentangled GAN Controls

Perla Doubinsky[a,**], Nicolas Audebert[a], Michel Crucianu[a], Hervé Le Borgne[b]

[a]*CEDRIC (EA4329), Conservatoire national des arts et métiers, Paris 75003, France*
[b]*Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France*

Article history:

GANs, image editing, latent space

ABSTRACT

Various controls over the generated data can be extracted from the latent space of a pre-trained GAN, as it implicitly encodes the semantics of the training data. The discovered controls allow to vary semantic attributes in the generated images but usually lead to entangled edits that affect multiple attributes at the same time. Supervised approaches typically sample and annotate a collection of latent codes, then train classifiers in the latent space to identify the controls. Since the data generated by GANs reflects the biases of the original dataset, so do the resulting semantic controls. We propose to address disentanglement by balancing the semantics of the dataset before training the classifiers. We demonstrate the effectiveness of this approach by extracting disentangled linear directions for face manipulation on state-of-the-art GAN architectures (including StyleGAN2 and StyleGAN3) and two datasets, CelebAHQ and FFHQ. We show that this simple and general approach outperforms state-of-the-art classifier-based methods while avoiding the need for disentanglement-enforcing post-processing.

1. Introduction

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) produce high-resolution and photorealistic images by learning a mapping between a latent space, modelled by a random distribution, and the real image space. New images can then easily be obtained by randomly sampling in the latent space and feeding the latent codes to the generator. However, their semantic properties might not be the desired ones. In applications such as data augmentation, it could be desirable to finely control the semantic properties of a generated image, especially to synthesize images that are difficult to capture in practice.

Recent research aim at leveraging pre-trained unconditional GANs and exploring their latent space to uncover the controls they can provide over the generated data. In particular, some methods find linear directions that can be interpreted as variations of some semantic attributes across the latent space (Härkönen et al., 2020; Plumerault et al., 2020; Jahanian et al., 2020; Shen et al., 2020; Zhuang et al., 2021; Yang et al., 2020; Shen and Zhou, 2021; Voynov and Babenko, 2020). However, the discovered directions often do not allow disentangled edits, affecting multiple attributes instead of solely altering the desired one. Learning-based supervised methods commonly rely on a three-stages pipeline that consists in sampling a set of latent codes, then labelling the latent codes from the corresponding images using pre-trained image classifiers and finally, extracting the directions. As GANs learn to approximate the training data distribution that carries different kinds of biases, the sampling stage leads to generating biased datasets that can, in turn, affect the semantic directions. The third stage is often performed by training a linear classifier to separate latent codes corresponding to images with a desired attribute (positive set) from those corresponding to images without the desired attribute (negative set). The direction controlling the attribute is then taken as the vector orthogonal to the classifier's decision boundary (Hutchinson et al., 2019; Shen et al., 2020; Yang et al., 2020). Existing correlations among attributes in the generated data may cause the positive and negative sets of a target attribute to be strongly imbalanced in respect to other attributes, thus biasing the direction

**Corresponding author:
e-mail: perla.doubinsky@lecnam.net (Perla Doubinsky),
nicolas.audebert@cnam.fr (Nicolas Audebert),
michel.crucianu@cnam.fr (Michel Crucianu),
herve.le-borgne@cea.fr (Hervé Le Borgne)

towards those attributes.

For GAN control, we can identify three datasets that typically carry biases: (1) the one used for training the GAN, (2) the one employed for training the image classifiers, and (3) the GAN-generated data used for finding the semantic directions. As shown in Fig. 1, the biases mainly come from the GAN training set (1). As GANs carry and amplify the bias (Zhao et al., 2018), the GAN-generated dataset (3) is also biased. Reducing GAN bias typically requires an access to the original dataset and GAN retraining. Instead, we propose to reduce the bias in the GAN-generated dataset directly. Specifically, after sampling and labelling the latent codes, we adjust the sampling of this data to balance the attribute joint distributions and remove correlations.

We apply our method in the latent space of GANs trained for face synthesis to identify semantic directions corresponding to facial attributes. We conduct experiments on different state-of-the-art GAN models: PGGAN (Karras et al., 2018) pre-trained on CelebAHQ (Liu et al., 2015), StyleGAN, StyleGAN2 and StyleGAN3 pre-trained on FFHQ (Karras et al., 2019, 2020, 2021). We provide a quantitative and qualitative comparison with the popular framework InterFaceGAN (Shen et al., 2020). We show that our approach leads to directions that are naturally disentangled whereas InterFaceGAN requires a post-processing step to reduce entanglement. Instead of relying on linear classifiers, we also propose to directly use the direction connecting class centroids, and show that it gives meaningful attribute controls for well-balanced data. The code is available online.[1]

2. Related work

Early works on GANs uncovered some level of semantic structure in the latent space *e.g.* by applying vector arithmetic on the latent codes (Radford et al., 2016). Subsequent works focused on finding global directions in latent space corresponding to specific factors of variation ranging from geometric transformations (*e.g.* position, scale) (Jahanian et al., 2020; Plumerault et al., 2020; Spingarn et al., 2021), memorability (Goetschalckx et al., 2019) to facial attributes (Shen et al., 2020; Shen et al., 2020; Härkönen et al., 2020; Voynov and Babenko, 2020; Spingarn et al., 2021; Zhuang et al., 2021; Shen and Zhou, 2021). By varying the latent codes towards those directions, the corresponding semantic properties of a generated image can be modified. Recent proposals argue that semantics distribute non-linearly and locally (Abdal et al., 2021; Hou et al., 2020; Wang et al., 2021) but such methods are more expensive as they require to compute a specific manipulation for each input.

Unsupervised methods. Some works attempt to find semantic directions with self-supervised learning (Voynov and Babenko, 2020), unsupervised approaches in latent space such as PCA (Härkönen et al., 2020), or by leveraging the internal representation of GANs to derive closed-form solutions (Shen and Zhou, 2021; Spingarn et al., 2021). However, since the semantics associated with each direction have to be manually identified afterwards, the discovery of the directions of interest is not guaranteed. In contrast, supervised methods aim to find directions corresponding to specific transformations *a priori*.

Supervised methods. These methods typically sample a large number of latent codes, then annotate the corresponding synthesized images with semantic labels using pre-trained image classifiers (Shen et al., 2020; Hutchinson et al., 2019; Yang et al., 2020; Wang et al., 2021; Hou et al., 2020; Abdal et al., 2021) to obtain a set of pairs (latent code, semantic labels). This set can be employed to train linear classifiers and each semantic direction is defined as the normal vector to the classifier decision boundary (Hutchinson et al., 2019; Shen et al., 2020; Yang et al., 2020). The latent codes are sampled according to the latent space prior (usually a multivariate Gaussian), which transfers to the semantic directions the bias of the dataset used to train the generator. In contrast, we propose a subsampling method to obtain a collection of latent codes that is balanced w.r.t. multiple attributes and doesn't carry strong correlations, thus mitigating the propagation of bias.

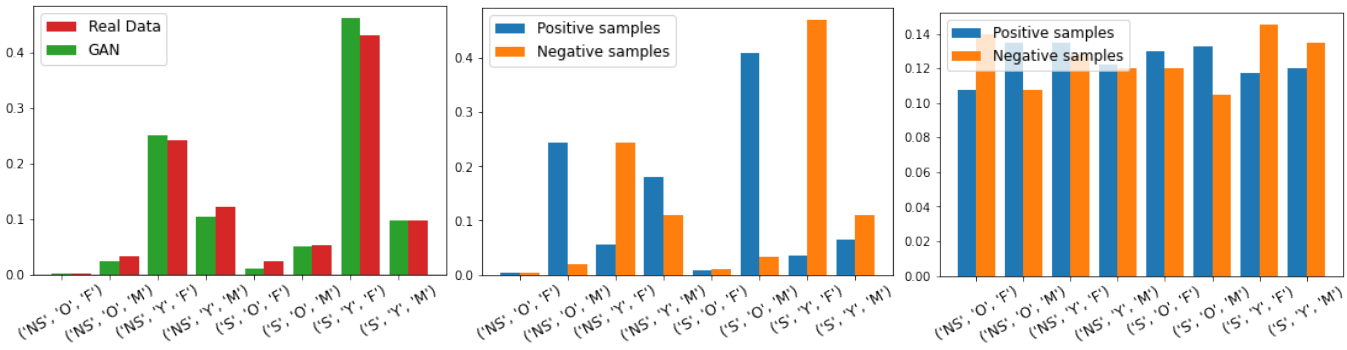
Disentanglement of semantics. Ideally, each of the discovered directions should control a single semantic property of the images. But very often the relation between directions and semantic properties is not one-to-one, *i.e.* one direction has an impact on several properties; one speaks of *entanglement*. To reduce entanglement, some propose to refine the semantic directions afterwards, by enforcing an orthogonality constraint for the new directions. This post-processing step is referred to as 'conditional manipulation' in (Shen et al., 2020; Wang et al., 2021). Spingarn *et al.* (Spingarn et al., 2021) introduce more constrained nonlinear paths that are defined as small circles on a sphere. Other works argue that entanglement is reduced if the transformations are learned together (Zhuang et al., 2021; Abdal et al., 2021). For style-based GAN architectures, Hou *et al.* (Hou et al., 2020) propose to learn an attention mechanism to manipulate the latent code for a particular layer. Differently from previous work, our method addresses entanglement *a priori* by debiasing the data employed to discover the directions. Hence, we argue that it can be complementary to previous proposals.

3. Balanced sampling and direction estimation

Let us consider a pre-trained generator $G(.)$ that maps a latent code $\mathbf{z}$ sampled from a d-dimensional latent space $\mathcal{Z} \subseteq \mathbb{R}^d$ to an image $\mathbf{I} = G(\mathbf{z})$ in image space $\mathcal{I} \subseteq \mathbb{R}^{H \times W \times C}$. Suppose the images are described by a set of binary attributes $\mathcal{A} = \{a_j, 1 \leq j \leq m\}$. For each attribute a_j we aim to find a global linear direction in the latent space, defined by unit vector $\mathbf{u}_j \in \mathbb{R}^d$, that allows to modify attribute a_j, and *only* attribute a_j, in a generated image by translating the corresponding latent code $\mathbf{z}$ in that direction, $\mathbf{z}' = \mathbf{z} + \alpha \mathbf{u}_j$, $\alpha \in \mathbb{R}$ being the moving step.

To find the directions, the procedure put forward in (Shen et al., 2020; Yang et al., 2020) is: (i) train a multi-attribute image classifier $F_{\mathcal{I}}$ on the ground truth provided with the database (*e.g.* CelebA (Liu et al., 2015)); (ii) generate N latent codes and corresponding images $\{(\mathbf{z}_i, G(\mathbf{z}_i))_{i=1}^N\}$; (iii) label every image with the classifier and associate the labels to the latent codes to produce $\mathcal{S} = \{(\mathbf{z}_i, F_{\mathcal{I}}(G(\mathbf{z}_i)))_{i=1}^N\}$; (iv) for each attribute j, train a linear classifier Ψ_j in latent space on the $\mathcal{S}_j^+$ and $\mathcal{S}_j^-$ sets obtained from

[1] Code : https://github.com/perladoubinsky/balanced_sampling_gan_controls.



(a) CelebAHQ vs. generated data with PGGAN CelebAHQ.

(b) Positives vs. negatives w.r.t. Glasses for random sampling.

(c) Positives vs. negatives w.r.t. Glasses for our sampling.

Fig. 1: Joint distributions for three binary facial attributes Age ('O': Old, 'Y': Young), Gender ('M': Male, 'F': Female) and Smile ('S': Smile, 'NS': No Smile). In (b), the positive set contains a majority of *old males* while the negative set contains a majority of *young females*, leading to bias the direction 'glasses' toward the attributes 'age' and 'gender'.

$\mathcal{S}$ by only considering the positive and respectively negative labels for attribute j. The direction in latent space allowing to control attribute j is then defined by $\mathbf{u}_j$ the unit vector that is orthogonal to the decision boundary of the linear classifier Ψ_j.

3.1. Multi-attribute balanced sampling

The distribution of the binary attributes for a set of data can be represented in an m-dimensional contingency table (one dimension per attribute) where each of the 2^m cells contains the number of samples that have the corresponding combination of values for the m attributes. If there are strong correlations between attributes in the GAN training data then the contingency table for that data is strongly imbalanced. The data in $\mathcal{S}$, generated by the trained GAN, is expected to show similar correlations. The example in Fig. 1 (a) reveals that three attributes in the CelebA (Liu et al., 2015) dataset are strongly correlated (some combinations are much more frequent than others) and this reflects well in the random sample generated by the GAN[2]. For an attribute a_j, the sets $\mathcal{S}_j^+$ and $\mathcal{S}_j^-$ employed for training a classifier in the latent space mirror the imbalance in $\mathcal{S}$. If we consider the attribute 'Glasses' in CelebA, Fig. 1 (b) shows how imbalanced the associated $\mathcal{S}_j^+$ and $\mathcal{S}_j^-$ sets are with respect to the three attributes in Fig. 1 (a). It is natural to expect that the classifier Ψ_j trained on such imbalanced data is influenced by the strong correlations. And, consequently, the unit vector $\mathbf{u}_j$ that is orthogonal to its decision boundary entangles the control of the target attribute with the most correlated attributes.

The idea of the method we propose is simple: subsample the data in $\mathcal{S}$ so as to obtain approximately the same number of samples in each cell of the contingency table. By removing the correlations,we expect to strongly reduce the entanglement.

More precisely, we build a multi-attribute balanced sample $\mathcal{B} \subset \mathcal{S}$ by iteratively selecting data from $\mathcal{S}$ until we reach the total number of samples $N_0 \leq N$ we aim to obtain. At each iteration, we first uniformly sample one combination of attribute values (one cell of the contingency table), then we uniformly sample without replacement one data point $(\mathbf{z}, F_{\mathcal{I}}(G(\mathbf{z})))$ with that combination. In this way, at the end of the sampling procedure, we expect to have a balanced contingency table for $\mathcal{B}$ where each of the 2^m cells contains approximately $\frac{N_0}{2^m}$ data points, as shown in Fig. 1 (c). The procedure is outlined in Algorithm 1.

The subsampling procedure works well if there is enough data in $\mathcal{S}$ for each combination of attribute values. For strongly imbalanced data, we may have to address the case where there is no more data in $\mathcal{S}$ for one or more combinations before reaching the desired total number of samples N_0. Note that, as we show in Section 4, good results can be obtained with moderate values for N_0. The ideal solution for having a balanced $\mathcal{B}$ is to expand $\mathcal{S}$ by generating more images with G. But this can be very expensive since, as we found, the imbalance of $\mathcal{S}$ reflects the imbalance of the training dataset. Hence, we may require the generation of a very large number of images to obtain one more image with a rare combination of attribute values.

Instead, the solution we adopt consists in simply skipping the current iteration if no more data is available for that combination. For high values of N_0, the resulting $\mathcal{B}$ is no longer so well-balanced, as we show in Section 4.2, this causes a slight decay in performance. An alternative is to oversample the already generated data corresponding to the rarest combinations of attribute values, *i.e.* random sample *with* replacement for a combination if its cell in the contingency table of $\mathcal{S}$ has much less than $\frac{N_0}{2^m}$ data points. As shown in Section 4.2, this allows to maintain good performances for high values of N_0.

3.2. Direction estimation

The sampling procedure we described leads to a sample $\mathcal{B}$ of size N_0 that is balanced w.r.t. all attributes. For each attribute j, two sets $\mathcal{B}_j^+$ of size $N_j^+ \approx \frac{N_0}{2}$ and $\mathcal{B}_j^-$ of size $N_j^- \approx \frac{N_0}{2}$ can be readily obtained by considering the data having positive and respectively negative labels for attribute j. To find the direction $\mathbf{u}_j$ in latent space that allows to control attribute j, a good solution is to train a linear classifier on $\mathcal{B}_j^+ \cup \mathcal{B}_j^-$, then take as $\mathbf{u}_j$ the vector orthogonal to the decision boundary. Preference is usually given (*e.g.* (Shen et al., 2020)) to linear Support Vector Machines (SVMs) that are fast to train and effective in high dimensions. To improve generalization, the value of the regularization hyperparameter could be selected by cross-validation. But as we

[2]Other attributes in CelebA are also strongly correlated.

Data: $\mathcal{S}$ a list of N labeled latent codes, $\mathcal{A}$ the corresponding multi-attribute labels, N_0 target number of samples
Result: N_0 latent codes balanced over $\mathcal{A}$
for *every attribute combination* $(a_1, a_2, \ldots, a_m) \in \mathcal{A}$ **do**
  $C[a_1, a_2, \ldots, a_m] \leftarrow$ latent codes of $\mathcal{S}$ labeled with this set of attributes;
end
$\mathcal{B} \leftarrow [\,]$;
for $i \leftarrow 1...N_0$ **do**
  $a_1, a_2, \ldots, a_m \leftarrow$ a random non-empty cell of C;
  $s \leftarrow$ a random latent code from $C[a_1, a_2, \ldots, a_m]$;
  remove s from C;
  $\mathcal{B} \leftarrow \mathcal{B} \cup s$;
end
return $\mathcal{B}$

Algorithm 1: Multi-attribute balanced sampling.

find later in Section 4.3, when the dataset is balanced, a stronger regularization (larger SVM margin) tends to produce directions that allow more disentangled edits. If the linear SVM has a very large margin, the decision boundary becomes orthogonal to the line connecting the centroids of the two classes. For attribute a_j, this direction is defined by:

$$\mathbf{u}_j = \frac{1}{N_j^+}\sum_{i=1}^{N_j^+}\mathbf{z}_i^+ - \frac{1}{N_j^-}\sum_{i=1}^{N_j^-}\mathbf{z}_i^-, \quad \mathbf{z}^+ \in B_j^+ \text{ and } \mathbf{z}^- \in B_j^-. \tag{1}$$

Experiments in Section 4.3 show that entanglement is further reduced when using this easy-to-compute direction.

4. Experiments

We compare our proposal with the state-of-the-art method InterFaceGAN (Shen et al., 2020). The main attributes we consider are 'glasses', 'gender', 'smile' and 'age'. The corresponding attribute control directions respectively produce the following effects: wearing glasses, presenting as male, smiling and getting younger. Section 4.1 provides a detailed quantitative analysis regarding entanglement. The impact of using a larger sample size is evaluated in Section 4.2, while in Section 4.3 we study the influence of the SVM regularization parameter.

Models. We conduct experiments with state-of-the-art GAN models trained on two face datasets, PGGAN trained on CelebAHQ (Karras et al., 2018) and StyleGAN, StyleGAN2 and StyleGAN3 trained on FFHQ (Karras et al., 2019, 2020, 2021). All models generate 1024×1024 images. Following (Shen et al., 2020), we train an auxiliary classifier on CelebA (Liu et al., 2015) with a ResNet-50 (He et al.) using multi-task learning to predict the attributes simultaneously. For each attribute, the task is a bi-classification problem with a softmax cross-entropy loss. We ensure that the accuracy of the classifier is above 80%.

Implementations details. We synthesize $N = 1M$ images with PGGAN and $N = 500K$ images with StyleGAN models. We prepare a larger dataset for PGGAN as some combinations of attributes are rarer in CelebAHQ than in FFHQ. We apply the attribute predictors to all the generated images and discard the samples having a confidence below 0.9. For each attribute, we collect $N_0 = 1000$ samples using our multi-attribute balanced sampling. We choose this value depending on the number of samples in the cell with fewest samples. The semantic directions are then obtained by taking the direction defined by the centroids of each class (see Section 3.2). For a fair comparison, we reproduce InterFaceGAN results instead of using the provided directions as they were not computed using the same attribute prediction model[3] nor the same number of samples. For InterFaceGAN, we uniformly subsample the generated dataset then train linear SVMs with $C = 1.0$ [4] to obtain the semantic directions given by unit vectors. These vectors are $512d$ (dimension of the latent spaces of PGGAN and StyleGAN).

Table 1: Re-scoring results for PGGAN. For each method, $\Delta\mathbf{e}$: overall entanglement, $\Delta\mathbf{r}$: effect. We highlight the best results among the disentangling approaches (IfGAN$^\perp$, Ours).

		Glasses	Gender	Smile	Age
IfGAN	$\Delta\mathbf{e}\downarrow$	0.205	0.118	0.034	0.125
	$\Delta\mathbf{r}\uparrow$	*0.386*	*0.519*	*0.386*	*0.142*
IfGAN$^\perp$	$\Delta\mathbf{e}\downarrow$	0.055	**0.018**	0.015	**0.055**
	$\Delta\mathbf{r}\uparrow$	*0.231*	*0.420*	*0.381*	*0.115*
Ours	$\Delta\mathbf{e}\downarrow$	**0.038**	0.041	**0.013**	0.072
	$\Delta\mathbf{r}\uparrow$	*0.286*	*0.448*	*0.370*	*0.129*

Metric. As in Shen et al. (2020), we use the re-scoring metric to quantify the desired effect and entanglement associated with a direction. This metric measures how the attribute scores vary after manipulating the latent codes. Intuitively, a good direction should induce an increase in the score corresponding to the target attribute while not affecting other scores. Given a direction $\mathbf{u}_j$ for attribute a_j, the re-scoring for attribute a_k is computed as:

$$\Delta\mathbf{s}_k = \frac{1}{n}\sum_{i=1}^{n}\left[F_{\mathcal{I},k}(G(\mathbf{z}_i)) - F_{\mathcal{I},k}(G(\mathbf{z}_i + \alpha\mathbf{u}_j))\right] \tag{2}$$

The desired *effect* $\Delta\mathbf{r}$ of direction $\mathbf{u}_j$ is given by the re-scoring result for the target attribute a_j. The *entanglement* of direction $\mathbf{u}_j$ with another attribute a_k is given by the re-scoring for that attribute. We derive a metric based on re-scoring to obtain the *overall entanglement* $\Delta\mathbf{e}$ associated with a direction. Similarly to StyleSpace Wu et al. (2021), we average the re-scoring results over the non-target attributes: $\Delta\mathbf{e} = \frac{1}{|\mathcal{A}|-1}\sum_{i\in\mathcal{A}\setminus a_j}|\Delta\mathbf{s}_i|$.

Re-scoring is evaluated on $n = 2000$ latent codes with $\alpha = 2.0$ for the editing and averaged over 3 experiments.

4.1. Disentanglement analysis

PGGAN is a traditional GAN architecture where a code is sampled from a Gaussian latent space $\mathcal{Z}$ and fed to the first convolutional layer. In addition to $\mathcal{Z}$, StyleGAN introduces an intermediate latent space $\mathcal{W}$ whose distribution is modelled by

[3] The model was not made available by the authors.

[4] As in the code provided by the authors: `https://github.com/genforce/interfacegan`.

Table 2: Re-scoring results for StyleGAN models in $\mathcal{Z}$ (top) and $\mathcal{W}$ (bottom). For each method, $\Delta\mathbf{e}$: overall entanglement, $\Delta\mathbf{r}$: effect. In $\mathcal{Z}$, we highlight the best results among the disentangling approaches (IfGAN$^{\perp}$, Ours).

(a) StyleGAN $\mathcal{Z}$

		Glasses	Gender	Smile	Age
IfGAN	$\Delta\mathbf{e}\downarrow$	0.140	0.161	0.050	0.108
	$\Delta\mathbf{r}\uparrow$	*0.339*	*0.335*	*0.154*	*0.156*
IfGAN$^{\perp}$	$\Delta\mathbf{e}\downarrow$	0.064	0.061	0.033	0.060
	$\Delta\mathbf{r}\uparrow$	*0.278*	*0.266*	*0.145*	*0.131*
Ours	$\Delta\mathbf{e}\downarrow$	**0.042**	**0.060**	**0.024**	**0.054**
	$\Delta\mathbf{r}\uparrow$	*0.345*	*0.307*	*0.173*	*0.142*

(b) StyleGAN2 $\mathcal{Z}$

		Glasses	Gender	Smile	Age
IfGAN	$\Delta\mathbf{e}\downarrow$	0.135	0.106	0.041	0.111
	$\Delta\mathbf{r}\uparrow$	*0.335*	*0.406*	*0.100*	*0.113*
IfGAN$^{\perp}$	$\Delta\mathbf{e}\downarrow$	0.049	**0.047**	0.026	0.068
	$\Delta\mathbf{r}\uparrow$	*0.232*	*0.346*	*0.099*	*0.091*
Ours	$\Delta\mathbf{e}\downarrow$	**0.038**	0.059	**0.024**	**0.062**
	$\Delta\mathbf{r}\uparrow$	*0.290*	*0.386*	*0.111*	*0.097*

(c) StyleGAN3 $\mathcal{Z}$

		Glasses	Gender	Smile	Age
IfGAN	$\Delta\mathbf{e}\downarrow$	0.121	0.122	0.041	0.153
	$\Delta\mathbf{r}\uparrow$	*0.444*	*0.395*	*0.301*	*0.182*
IfGAN$^{\perp}$	$\Delta\mathbf{e}\downarrow$	**0.030**	0.037	0.018	**0.069**
	$\Delta\mathbf{r}\uparrow$	*0.382*	*0.346*	*0.295*	*0.163*
Ours	$\Delta\mathbf{e}\downarrow$	0.044	**0.027**	**0.014**	0.076
	$\Delta\mathbf{r}\uparrow$	*0.398*	*0.377*	*0.305*	*0.172*

(d) StyleGAN $\mathcal{W}$

		Glasses	Gender	Smile	Age
IfGAN	$\Delta\mathbf{e}\downarrow$	0.046	0.140	0.073	0.076
	$\Delta\mathbf{r}\uparrow$	*0.480*	*0.370*	*0.237*	*0.167*
Ours	$\Delta\mathbf{e}\downarrow$	**0.033**	**0.073**	**0.052**	**0.046**
	$\Delta\mathbf{r}\uparrow$	*0.603*	*0.435*	*0.238*	*0.169*

(e) StyleGAN2 $\mathcal{W}$

		Glasses	Gender	Smile	Age
IfGAN	$\Delta\mathbf{e}\downarrow$	0.040	**0.030**	**0.025**	**0.021**
	$\Delta\mathbf{r}\uparrow$	*0.207*	*0.245*	*0.132*	*0.092*
Ours	$\Delta\mathbf{e}\downarrow$	**0.031**	0.031	0.026	0.055
	$\Delta\mathbf{r}\uparrow$	*0.278*	*0.294*	*0.129*	*0.102*

(f) StyleGAN3 $\mathcal{W}$

		Glasses	Gender	Smile	Age
IfGAN	$\Delta\mathbf{e}\downarrow$	0.065	0.062	0.024	0.077
	$\Delta\mathbf{r}\uparrow$	*0.417*	*0.229*	*0.248*	*0.145*
Ours	$\Delta\mathbf{e}\downarrow$	**0.036**	**0.037**	**0.009**	**0.051**
	$\Delta\mathbf{r}\uparrow$	*0.383*	*0.303*	*0.287*	*0.146*

Fig. 2: Editing results for PGGAN for attributes Glasses, Gender and Age.

Table 3: Re-scoring results for StyleGAN3 in $\mathcal{Z}$ and $\mathcal{W}$ for rare attributes. For each method, $\Delta\mathbf{e}$: overall entanglement, $\Delta\mathbf{r}$: effect. In $\mathcal{Z}$, we highlight the best results among the disentangling approaches (IfGAN$^{\perp}$, Ours).

(a) StyleGAN3 $\mathcal{Z}$

		Pale Skin [a]	Wavy Hair [a]	Narrow Eyes [a]
IfGAN	$\Delta\mathbf{e}\downarrow$	0.106	0.180	0.165
	$\Delta\mathbf{r}\uparrow$	*0.277*	*0.302*	*0.234*
IfGAN$^{\perp}$	$\Delta\mathbf{e}\downarrow$	0.036	**0.068**	0.096
	$\Delta\mathbf{r}\uparrow$	*0.251*	*0.273*	*0.208*
Ours	$\Delta\mathbf{e}\downarrow$	**0.027**	0.083	**0.059**
	$\Delta\mathbf{r}\uparrow$	*0.308*	*0.341*	*0.239*

[a] Pale skin is balanced w.r.t. Gender and Age, Wavy Hair w.r.t. to Gender and Narrow Eyes w.r.t. Smile.

(b) StyleGAN3 $\mathcal{W}$

		Pale Skin	Wavy Hair	Narrow Eyes
IfGAN	$\Delta\mathbf{e}\downarrow$	0.038	0.059	0.120
	$\Delta\mathbf{r}\uparrow$	*0.256*	*0.153*	*0.196*
Ours	$\Delta\mathbf{e}\downarrow$	**0.035**	**0.032**	**0.055**
	$\Delta\mathbf{r}\uparrow$	*0.252*	*0.260*	*0.245*

fully-connected layers and learned during training, leading to a less entangled space (Karras et al., 2019). We compare our method to InterFaceGAN before (IfGAN) and after conditional manipulation (IfGAN$^{\perp}$), the latter having been introduced as an *ad hoc* disentanglement post-processing (Shen et al., 2020). For attribute j, it replaces $\mathbf{u}_j$ by its projection on the subspace orthogonal to the directions found for the other attributes.

PGGAN. The results in Table 1 show a strong entanglement for IfGAN especially for the attributes 'glasses', 'gender' and 'age', which are the most correlated attributes. These results reflect what can be observed in Fig. 2 for instance: the direction 'gender' tends to age the face and the direction 'age' tends to feminize the face. The conditional manipulation allows to reduce the entanglement while maintaining the desired effect. Our approach succeeds to extract directions allowing disentangled edits without requiring conditional manipulation. It significantly outperforms IfGAN and performs on par with IfGAN$^{\perp}$.

We experimented with applying the conditional manipulation to post-process our directions. However, we found that they are initially sufficiently orthogonal and that enforcing this geometrical constraint actually slightly increases the entanglement.

StyleGAN models. Tables 2a to 2c show there is less entanglement in $\mathcal{Z}$ space of the StyleGAN models compared to PGGAN, probably because FFHQ is a larger dataset and the attributes are less correlated than in CelebAHQ. Otherwise, we find similar tendencies. The $\mathcal{W}$ space being less entangled than $\mathcal{Z}$, the results for IfGAN shown in Tables 2d to 2f are good (conditional manipulation is not necessary). Fig. 3 also shows better editing results in $\mathcal{W}$ space as evident for the attribute

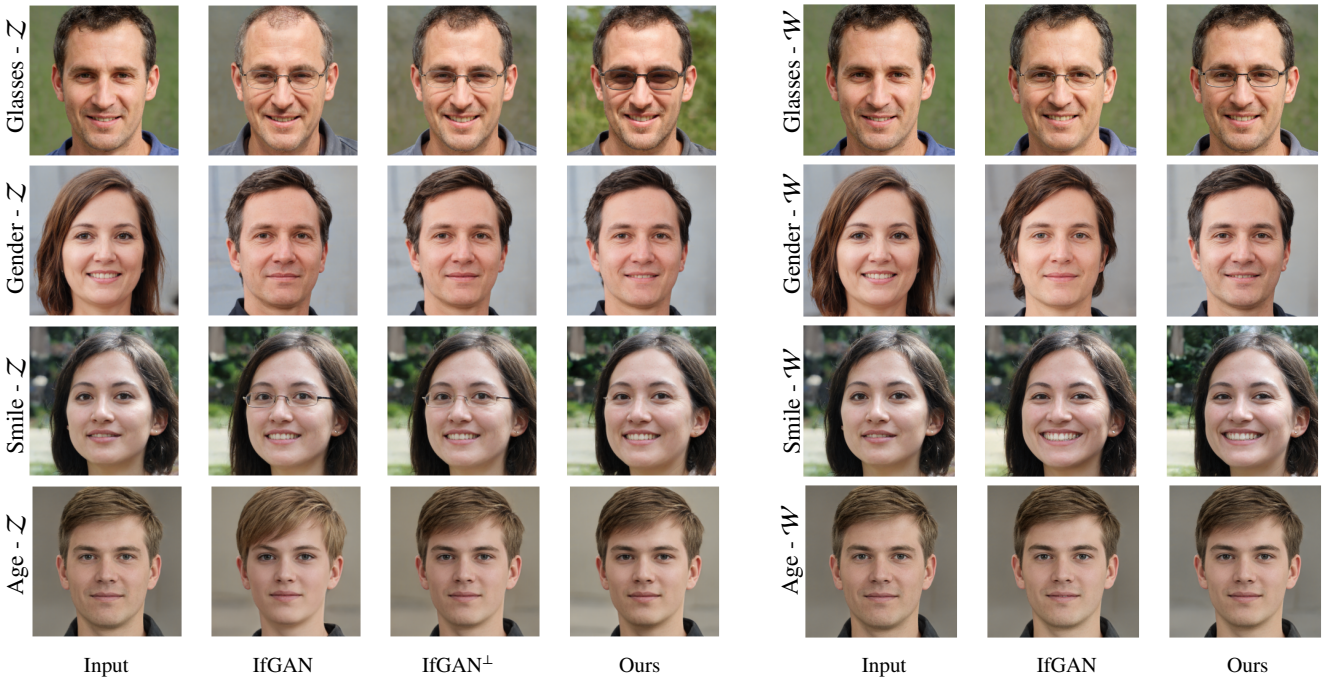


Fig. 3: Editing results for StyleGAN2 for attributes Glasses, Gender, Smile and Age.

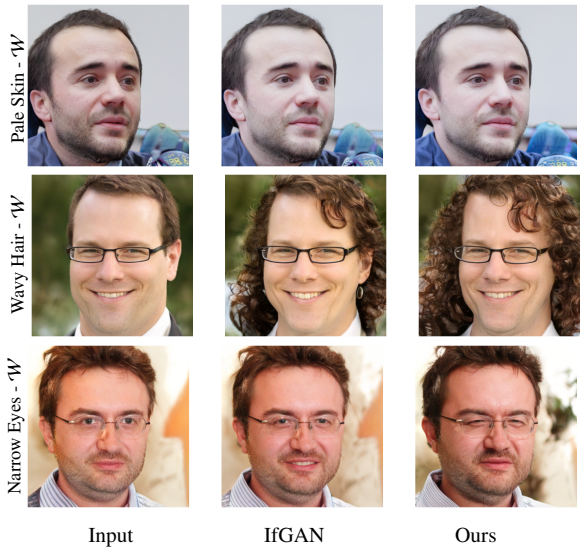


Fig. 4: Editing results for StyleGAN3 in $\mathcal{W}$ for attributes Pale Skin, Wavy Hair and Narrow Eyes.

'glasses'. Our method reaches similar to slightly better results than IfGAN, except for the attribute 'age' in StyleGAN2.

Table 3 presents results for different attributes: 'pale skin', 'wavy hair' and 'narrow eyes' for the StyleGAN3 model. Despite being quite rare, we can find effective directions for these attributes. In particular, we notice that the attribute 'narrow eyes' is not naturally disentangled from 'smile' in $\mathcal{W}$ space. Our method allows to significantly reduce the entanglement with this attribute as shown in Fig. 4. Fig. 5 shows additional qualitative results with intermediate α values for StyleGAN3 in $\mathcal{W}$ space.

Similarly to IfGAN, we also investigate the preservation of identity (cosine similarity between VGG-Face Cao et al. (2018) features before and after manipulation). However, this metric is not relevant for all attributes (*e.g.*'gender') and is dependent on the effect of a given direction. For suitable attributes (*e.g.*'smile') and in relation to the effect, we obtain similar results to IfGAN.

4.2. Impact of the sample size

We study the influence of using a larger sample size to estimate the directions. The directions are calculated with a sample of size $N_0 = 10000$ (instead of $N_0 = 1000$). For a larger sample size, the distributions are no longer well-balanced as potentially many cells of the contingency tables have been emptied. As shown in Table 4, the entanglement for the attribute 'glasses' remains quite low but the entanglement for the attribute 'gender' and 'age' is quite high, almost similar to IfGAN. To mitigate this effect, we propose to oversample the rarest samples as mentioned in Section 3.1. The results (ours* in Table 4) show that it allows to decrease the entanglement and to finally reach a similar entanglement as for the directions estimated with a sample of size $N_0 = 1000$. We also find that the directions seem to have more effect when using a larger sample size. Following these observations, we argue that a large sample size (as in Shen et al. (2020)) is not necessary to obtain meaningful directions. Nevertheless, oversampling allows to increase sample size for a stronger effect, while keeping a low entanglement.

4.3. SVM vs. centroids difference

The calculation of the directions is usually performed using SVMs trained in latent space. We study the influence of the regularization parameter on the extracted directions. Table 5 shows that a stronger regularization (lower C) leads to smaller entanglement, while the effect on the target attribute remains almost unchanged. This observation led us to consider the case of a very large SVM margin, when the decision boundary becomes orthogonal to the direction connecting the centroids of the two classes (see Section 3.2). This direction performs best.

Table 4: Re-scoring results for a higher sample size $N_0 = 10K$ for different models: PGGAN, StyleGAN3 in $\mathcal{Z}$ and in $\mathcal{W}$, $\Delta\mathbf{e}$: overall entanglement, $\Delta\mathbf{r}$: effect.

		(a) PGGAN				(b) StyleGAN3 $\mathcal{Z}$				(c) StyleGAN3 $\mathcal{W}$			
		Glasses	Gender	Smile	Age	Glasses	Gender	Smile	Age	Glasses	Gender	Smile	Age
IfGAN $N_0 = 10K$	$\Delta\mathbf{e}\downarrow$	0.238	0.141	0.034	0.137	0.167	0.196	0.043	0.183	0.078	0.087	0.030	0.075
	$\Delta\mathbf{r}\uparrow$	*0.488*	*0.560*	*0.411*	*0.149*	*0.570*	*0.464*	*0.346*	*0.192*	*0.527*	*0.305*	*0.292*	*0.152*
Ours $N_0 = 10K$	$\Delta\mathbf{e}\downarrow$	0.099	0.099	0.017	0.108	**0.040**	0.114	0.042	0.123	0.037	0.066	0.033	0.072
	$\Delta\mathbf{r}\uparrow$	*0.415*	*0.543*	*0.407*	*0.140*	*0.523*	*0.444*	*0.341*	*0.184*	*0.414*	*0.321*	*0.299*	*0.152*
Ours* $N_0 = 10K$	$\Delta\mathbf{e}\downarrow$	**0.055**	**0.035**	**0.003**	**0.078**	0.044	**0.048**	**0.018**	**0.087**	**0.036**	**0.031**	**0.009**	**0.051**
	$\Delta\mathbf{r}\uparrow$	*0.367*	*0.505*	*0.400*	*0.136*	*0.515*	*0.419*	*0.335*	*0.182*	*0.413*	*0.309*	*0.296*	*0.150*
Ours $N_0 = 1K$	$\Delta\mathbf{e}\downarrow$	0.038	0.041	0.013	0.072	0.044	0.027	0.014	0.076	0.036	0.037	0.009	0.051
	$\Delta\mathbf{r}\uparrow$	*0.286*	*0.448*	*0.370*	*0.129*	*0.398*	*0.377*	*0.305*	*0.172*	*0.383*	*0.303*	*0.287*	*0.146*

Table 5: Re-scoring results for different boundary calculation methods (given a balanced sample) for different models: PGGAN, StyleGAN3 in $\mathcal{Z}$ and $\mathcal{W}$, $\Delta\mathbf{e}$: overall entanglement, $\Delta\mathbf{r}$: effect.

		(a) PGGAN				(b) StyleGAN3 $\mathcal{Z}$				(c) StyleGAN3 $\mathcal{W}$			
		Glasses	Gender	Smile	Age	Glasses	Gender	Smile	Age	Glasses	Gender	Smile	Age
SVM $C = 1.0$	$\Delta\mathbf{e}\downarrow$	0.118	0.080	0.011	0.113	0.069	0.108	0.060	0.114	0.051	0.054	0.021	0.054
	$\Delta\mathbf{r}\uparrow$	*0.326*	*0.500*	*0.382*	*0.136*	*0.423*	*0.3901*	*0.296*	*0.177*	*0.404*	*0.232*	*0.261*	*0.144*
SVM $C = 0.001$	$\Delta\mathbf{e}\downarrow$	0.069	0.045	**0.010**	0.091	0.054	0.067	0.029	0.087	0.042	0.044	0.019	0.055
	$\Delta\mathbf{r}\uparrow$	*0.323*	*0.471*	*0.379*	*0.133*	*0.431*	*0.391*	*0.310*	*0.176*	*0.409*	*0.296*	*0.291*	*0.155*
centroids	$\Delta\mathbf{e}\downarrow$	**0.038**	**0.041**	0.013	**0.072**	**0.044**	**0.027**	**0.014**	**0.076**	**0.036**	**0.037**	**0.009**	**0.051**
	$\Delta\mathbf{r}\uparrow$	*0.286*	*0.448*	*0.370*	*0.129*	*0.398*	*0.377*	*0.305*	*0.172*	*0.383*	*0.303*	*0.287*	*0.146*

5. Discussion

While our method balances the sample to decorrelate the attributes, we observed that the resulting directions in latent space are quasi-orthogonal, which was not *a priori* expected. This may explain the success of previous works that look for orthogonal directions in the latent space. For example, GANSpace (Härkönen et al., 2020) applies PCA in the $\mathcal{W}$ space and the authors are able to assign semantic interpretations to the resulting directions (orthogonal by definition). The conditional manipulation in InterFaceGAN (Shen et al., 2020) also enforces an orthogonality constraint among control directions to reduce entanglement. This requirement of orthogonality did not have an *a priori* justification but our results indicate that orthogonality in latent space could be a necessary condition for independent controls and, even for unconditional GANs, the latent space does encode a significant part of the semantics. We believe that our subsampling approach can prove beneficial to other works on GAN control that rely on sampling in the latent space. Three issues could be raised. First, as in most works on finding supervised controls, we use pseudo-labels provided by image classifiers that are assumed reliable. But they can also be affected by bias, with an impact on both the labelling of the training set and the evaluation since re-scoring depends on the classifiers. However, results on FFHQ show that even classifiers trained on smaller datasets like CelebAHQ transfer quite well. Second, using classifiers to find directions assumes that samples can be grouped in classes. This nevertheless works surprisingly well for binarized continuous attributes (*e.g.*'age') and might not be a problem in practice.

Finally, our method only balances known attributes. Entanglements due to representation biases of unlabeled attributes can remain, and in rare occasions, be worsened by the oversampling of rare combinations. This underlines that the set of attributes should be chosen and labeled carefully to achieve fair and unbiased editing.

6. Conclusion

We focused on the identification of directions in the latent space of a GAN to control semantic attributes of the generated images. Our assumption was that the entanglement typically observed in such situations results from strong correlations among attributes in the training data, that are transferred to the generated data. To address this issue, we proposed a simple and general method that balances the data among the different combinations of values for the attributes. The evaluation on two popular GAN architectures and two face datasets shows that this approach outperforms state-of-the-art classifier-based methods while avoiding the need for post-processing. We believe it can prove useful to other sampling-based GAN control methods.

Acknowledgments

This research was performed under a grant from the AHEAD ANR program (ANR-20-THIA-0002) and supported by the European Commission under European Horizon 2020 Programme, grant number 951911 - AI4Media.

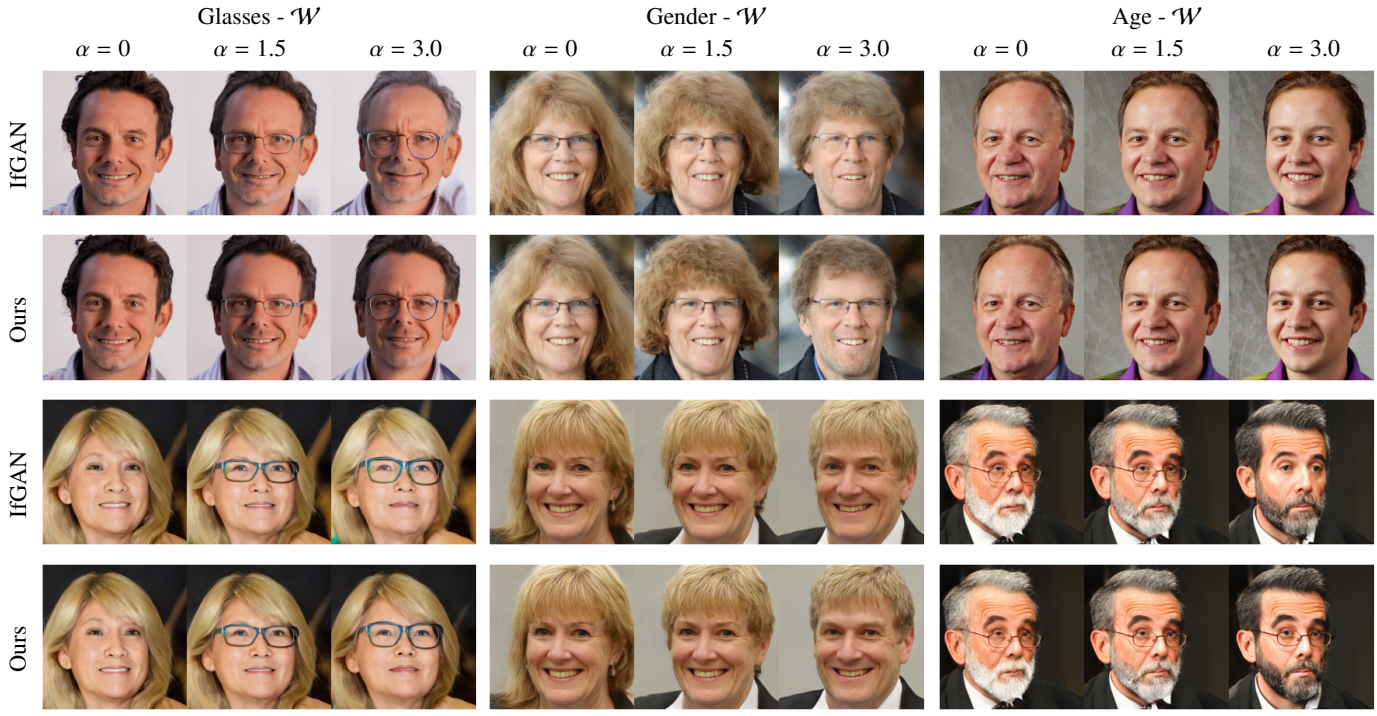


Fig. 5: Editing results for StyleGAN3 in $\mathcal{W}$ for attributes Glasses, Gender and Age.

References

Abdal, R., Zhu, P., Mitra, N.J., Wonka, P., 2021. StyleFlow: Attribute-conditioned exploration of StyleGAN-generated images using conditional continuous normalizing flows. ACM Transaction on Graphics 40.

Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A., 2018. VGGFace2: A dataset for recognising faces across pose and age, in: 2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018).

Goetschalckx, L., Andonian, A., Oliva, A., Isola, P., 2019. GANalyze: Toward visual definitions of cognitive image properties, in: Proceedings of the IEEE/CVF International Conference on Computer Vision.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets, in: Advances in Neural Information Processing Systems, Curran Associates, Inc.

He, K., Zhang, X., Ren, S., Sun, J., . Deep residual learning for image recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Hou, X., Zhang, X., Shen, L., Lai, Z., Wan, J., 2020. GuidedStyle: Attribute knowledge guided style manipulation for semantic face editing. ArXiv abs/2012.11856.

Hutchinson, B., Denton, E., Mitchell, M., Gebru, T., 2019. Detecting bias with generative counterfactual face attribute augmentation, in: Fairness, Accountability, Transparency and Ethics in Computer Vision Workshop.

Härkönen, E., Hertzmann, A., Lehtinen, J., Paris, S., 2020. GANSpace: Discovering interpretable gan controls, in: Advances in Neural Information Processing Systems.

Jahanian, A., Chai, L., Isola, P., 2020. On the "steerability" of generative adversarial networks, in: International Conference on Learning Representations.

Karras, T., Aila, T., Laine, S., Lehtinen, J., 2018. Progressive growing of GANs for improved quality, stability, and variation, in: International Conference on Learning Representations.

Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T., 2021. Alias-free generative adversarial networks, in: Proc. NeurIPS.

Karras, T., Laine, S., Aila, T., 2019. A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T., 2020. Analyzing and improving the image quality of StyleGAN, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Liu, Z., Luo, P., Wang, X., Tang, X., 2015. Deep learning face attributes in the wild, in: Proceedings of the IEEE/CVF International Conference on Computer Vision.

Plumerault, A., Le Borgne, H., Hudelot, C., 2020. Controlling generative models with continuous factors of variations, in: International Conference on Learning Representations.

Radford, A., Metz, L., Chintala, S., 2016. Unsupervised representation learning with deep convolutional generative adversarial networks, in: International Conference on Learning Representations.

Shen, Y., Gu, J., Tang, X., Zhou, B., 2020. Interpreting the Latent Space of GANs for Semantic Face Editing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Shen, Y., Yang, C., Tang, X., Zhou, B., 2020. InterFaceGAN: Interpreting the disentangled face representation learned by GANs. IEEE Transactions on Pattern Analysis and Machine Intelligence , 1–1.

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

Spingarn, N., Banner, R., Michaeli, T., 2021. GAN "steerability" without optimization, in: International Conference on Learning Representations.

Voynov, A., Babenko, A., 2020. Unsupervised discovery of interpretable directions in the GAN latent space, in: Proceedings of the 37th International Conference on Machine Learning, PMLR. pp. 9786–9796.

Wang, H.P., Yu, N., Fritz, M., 2021. Hijack-GAN: Unintended-use of pretrained, black-box GANs, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7872–7881.

Wu, Z., Lischinski, D., Shechtman, E., 2021. StyleSpace analysis: Disentangled controls for StyleGAN image generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Yang, C., Shen, Y., Zhou, B., 2020. Semantic hierarchy emerges in deep generative representations for scene synthesis. International Journal of Computer Vision .

Zhao, S., Ren, H., Yuan, A., Song, J., Goodman, N., Ermon, S., 2018. Bias and generalization in deep generative models: An empirical study. Advances in Neural Information Processing Systems 31.

Zhuang, P., Koyejo, O.O., Schwing, A., 2021. Enjoy your editing: Controllable GANs for image editing via latent space navigation, in: International Conference on Learning Representations.