Multi-Attribute Balanced Sampling for Disentangled GAN Controls

Perla Doubinsky, Nicolas Audebert, Michel Crucianu, Hervé Le Borgne

MULTI-ATTRIBUTE BALANCED SAMPLING FOR DISENTANGLED GAN CONTROLS

A PREPRINT

Perla Doubinsky
perla.doubinsky@lecnam.net

Nicolas Audebert
nicolas.audebert@cnam.fr

Michel Crucianu
michel.crucianu@cnam.fr

Hervé Le Borgne
herve.le-borgne@cea.fr

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 subsampling the generated data to remove over-represented co-occuring attributes thus 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 two popular GAN architectures, PGGAN and StyleGAN, and two datasets, CelebAHQ and FFHQ. We show that this approach outperforms state-of-the-art classifier-based methods while avoiding the need for disentanglement-enforcing post-processing.

1 Introduction

Generative Adversarial Networks (GANs) [1] 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 be obtained by randomly sampling in the latent space and feeding the latent codes to the generator. While it is easy to generate an image, its semantic properties might not be the desired ones. In applications such as data augmentation, it could be very useful 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 [2, 3, 4, 5, 6, 7, 8, 9]. 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 real 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 latent codes 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 [10, 5, 7]. 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 towards those attributes.

Inspired by this observation, we propose to enforce disentanglement by learning the semantic directions on datasets that are free from bias. Specifically, after sampling and labelling the latent codes, we subsample the dataset to balance the attributes 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 two types of GAN architectures: PGGAN [11] and StyleGAN [12], respectively pre-trained on CelebAHQ [13] and FFHQ [12]. We provide a quantitative and qualitative comparison with the popular framework InterFaceGAN [5]. We show that our approach leads to directions that are naturally disentangled whereas InterFaceGAN requires a post-processing step to reduce entanglement. We also show that, instead of relying on linear classifiers, directly using the direction connecting class centroids can give meaningful attribute controls for well-balanced data.

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 [14]. Subsequent works focused on finding global directions in latent space corresponding to specific factors of variation ranging from geometric transformations (*e.g.* position, scale) [4, 3, 15], memorability [16] to facial attributes [5, 17, 2, 9, 15, 6, 8]. 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 [18, 19, 20] 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 [9], unsupervised approaches in latent space such as PCA [2], or by leveraging the internal representation of GANs to derive closed-form solutions [8, 15]. 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*.

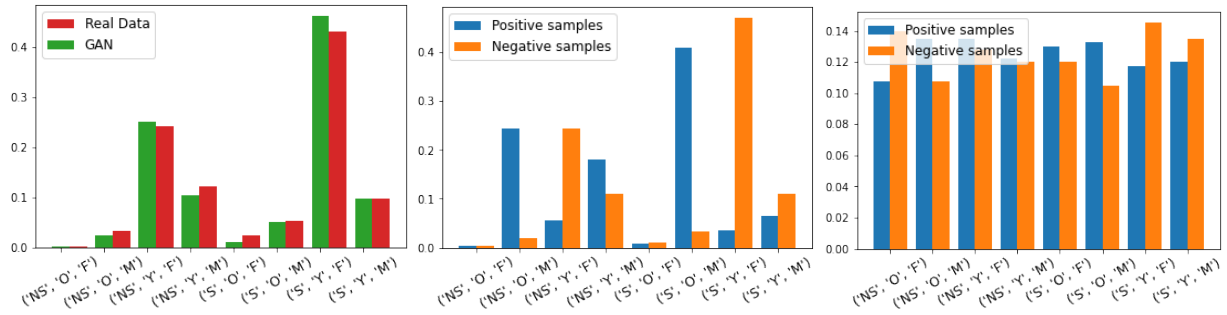
Supervised methods. These methods typically sample a large number of latent codes, then annotate the corresponding synthesized images with semantic labels using pre-trained image classifiers [5, 10, 7, 20, 19, 18] to obtain a set of pairs (latent code, semantic labels). This set can be employed to train linear classifiers and each semantic direction is defined as the normal vector to the classifier decision boundary [10, 5, 7]. 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 [5, 20]. Spingarn *et al*. [15] 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 [6, 18]. For style-based GAN architectures, Hou *et al*. [19] 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 [17, 7] is: (i) train a multi-attribute image classifier $F_{\mathcal{I}}$ on the ground truth provided with the database (*e.g.* CelebA [13]); (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 $\mathcal{S}$ by only considering the positive and respectively negative labels for attribute j. The direction in latent



(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.

Figure 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".

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 [13] dataset are strongly correlated (some combinations are much more frequent than others) and this reflects well in the random sample generated by the GAN[1]. 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 from this table, 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 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.

The solution we adopt consists in simply skipping the current iteration if no more data is available for that combination. The resulting $\mathcal{B}$ is no longer so well-balanced but, as we show in Section 4.2, there is a graceful decay in performance. An alternative is to oversample the data corresponding to the rarest combinations of attribute values, *i.e.* random sample

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

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 makes the decay in performance yet more graceful.

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.* [5]) 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 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 this easy-to-compute direction is used to control the corresponding attribute.

4 Experiments

We evaluate and compare our proposal with the state-of-the-art method InterFaceGAN [5], considering the same attributes "glasses", "gender", "smile" and "age". For InterFaceGAN, 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 of the effect a direction has for different attributes. The impact of sample size is evaluated in Section 4.2, while in Section 4.3 we study regularization. Identity preservation is assessed in Section 4.4.- Finally, qualitative results are shown in Section 4.5.

Models. We conduct experiments with state-of-the-art GAN models trained on two face datasets, PGGAN CelebAHQ [11] and StyleGAN FFHQ [12], generating 1024×1024 images (experiments with StyleGAN FFHQ trained on 256×256 images are shown in the supplementary material). Following [5], we train an auxiliary classifier on CelebA [13] with a ResNet-50 [21] 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 classifiers is above 80% (see details in the supplementary material).

Implementations details. We synthesize $N = 1M$ images with PGGAN CelebAHQ and $N = 500K$ images with StyleGAN FFHQ. 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 (contingency tables are given in the supplementary material). 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[2] nor the same number of samples. For InterFaceGAN, we uniformly subsample the generated dataset then train linear SVMs with $C = 1.0$ [3] to obtain the semantic directions given by unit vectors. Since the dimension of the latent spaces of PGGAN and StyleGAN is 512, these vectors are also $512d$.

Metrics. As in [5], 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 $\mathbf{u}_j$ corresponding to attribute a_j, the re-scoring for attribute a_k is computed as:

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

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

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

	Glasses	Gender	Smile	Age
Glasses	0.39	0.34	-0.06	-0.29
Gender	0.09	0.50	-0.06	-0.22
Smile	-0.03	-0.07	0.37	-0.02
Age	-0.04	-0.31	-0.07	0.15

(a) IfGAN [5]

	Glasses	Gender	Smile	Age
Glasses	0.27	0.10	-0.02	-0.10
Gender	-0.00	0.41	-0.03	-0.02
Smile	-0.02	-0.04	0.36	-0.01
Age	-0.02	-0.11	-0.04	0.13

(b) IfGAN + conditional [5]

	Glasses	Gender	Smile	Age
Glasses	0.32	0.07	-0.05	-0.07
Gender	0.01	0.42	-0.04	-0.05
Smile	-0.01	-0.02	0.36	-0.01
Age	-0.02	-0.13	-0.06	0.14

(c) Our method

Table 1: Re-scoring results in $\mathcal{Z}$ space for PGGAN CelebAHQ. Each row shows the effect of a direction on all attributes, each column shows the effect of all directions on an attribute. Effect (diagonal values) should be high, entanglement (off-diagonal values) should be low.

The desired *effect* of direction $\mathbf{u}_j$ is given by the re-scoring result for the target attribute a_j (higher is better). The *entanglement* of direction $\mathbf{u}_j$ with another attribute a_k is given by the re-scoring for that attribute (lower is better). We also derive a metric based on re-scoring to obtain the *overall entanglement* associated with a direction. Similarly to StyleSpace [22], we average the re-scoring results over the non-target attributes: $\frac{1}{|\mathcal{A}|-1}\sum_{i\in\mathcal{A}\setminus a_j}|\Delta\mathbf{s}_i|$. To quantify how the manipulations affect face identity, we employ a popular face recognition model pre-trained on VGGFace2 [23] and compute the cosine similarity between face embeddings before and after editing, as in [6]. We extract embeddings of dimension 2048.

Both metrics are evaluated on $n = 2000$ latent codes, we employ $\alpha = 0.2$ for the editing and we report averages over 5 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 fully-connected layers and learned during training, leading to a less entangled space [12]. We compare our method to InterFaceGAN before (IfGAN) and after conditional manipulation (IfGAN + conditional), the latter having been introduced as an *ad hoc* disentanglement post-processing [5]. For attribute j, it consists in replacing $\mathbf{u}_j$ by the its projection on the subspace orthogonal to the directions found for the other attributes.

PGGAN. Table 1 (a) provides the re-scoring results for InterFaceGAN without conditional manipulation. We observe that the diagonal scores increase after manipulating the latent codes, which shows that the directions have the desired effect. On the other hand, some of the off-diagonal scores also increase, indicating entanglement with other attributes. For instance, the direction "glasses" also affects the attributes "gender" and "age". As shown in Table 1 (b), the conditional manipulation allows to reduce the entanglement while maintaining the desired effect. According to Table 1 (c), our approach succeeds to extract directions allowing disentangled edits without requiring conditional manipulation. It outperforms significantly InterFaceGAN and performs slightly better than conditional manipulation.

StyleGAN. Table 2 shows the results in $\mathcal{Z}$ space. Compared to PGGAN there is less entanglement, probably because FFHQ is a larger dataset and the attributes are less correlated than in CelebAHQ. Otherwise, we find similar tendencies. In addition, the directions extracted with our approach have a stronger effect and are either on par or significantly more disentangled than for InterFaceGAN, even with conditional manipulation. The results in $\mathcal{W}$ space are given in Table 3. The $\mathcal{W}$ space being less entangled than $\mathcal{Z}$, the results of InterfaceGAN are good (conditional manipulation is not necessary) and those of our method are similar. At a lower resolution (StyleGAN 256×256), $\mathcal{W}$ is nevertheless less disentangled and our method significantly improves disentanglement w.r.t. InterfaceGAN (see supplementary material).

4.2 Impact of the sample size

We study the impact of the sample size on the effect of the extracted directions. For our method, we consider both settings, *i.e.* sampling without replacement and sampling with replacement (oversampling). Figure 2 shows the results in $\mathcal{Z}$ space for PGGAN CelebAHQ (similar results are given for StyleGAN FFHQ in the supplementary material, for $\mathcal{Z}$ and $\mathcal{W}$). For both our method and InterFaceGAN, we find that the effect and the entanglement increase with the size of the sample. For moderate values of N_0 the distributions are well-balanced, hence the level of entanglement of our directions remains significantly below that of InterFaceGAN directions. However, we observe a significant increase for $N_0 = 10\,000$, suggesting that the distributions are no longer balanced as many cells of the contingency table have been emptied (see tables in the supplementary material). We find that oversampling allows to mitigate this effect. Additionally, the size of the sample has a limited impact regarding direction variability between runs. Standard

	Glasses	Gender	Smile	Age
Glasses	0.36	0.23	-0.05	-0.19
Gender	0.16	0.37	-0.11	-0.18
Smile	0.03	-0.08	0.15	0.00
Age	-0.12	-0.25	0.00	0.18

(a) IfGAN [5]

	Glasses	Gender	Smile	Age
Glasses	0.27	0.12	-0.01	-0.08
Gender	0.06	0.29	-0.07	-0.07
Smile	-0.04	-0.06	0.15	0.00
Age	-0.07	-0.15	0.00	0.15

(b) IfGAN + conditional [5]

	Glasses	Gender	Smile	Age
Glasses	0.35	0.06	-0.02	-0.05
Gender	0.02	0.33	-0.08	-0.06
Smile	-0.01	-0.05	0.17	0.00
Age	-0.07	-0.13	-0.01	0.17

(c) Our method

Table 2: Re-scoring results in $\mathcal{Z}$ space for StyleGAN FFHQ.

	Glasses	Gender	Smile	Age
Glasses	0.51	0.09	0.02	-0.09
Gender	0.08	0.39	-0.22	-0.06
Smile	0.09	-0.14	0.22	-0.03
Age	-0.11	-0.11	-0.04	0.20

(a) IfGAN [5]

	Glasses	Gender	Smile	Age
Glasses	0.63	0.09	-0.07	-0.04
Gender	0.05	0.44	-0.10	-0.08
Smile	0.05	-0.06	0.22	-0.05
Age	-0.09	-0.15	-0.01	0.20

(b) Our method

Table 3: Re-scoring results in $\mathcal{W}$ space for StyleGAN FFHQ.

deviations increase when we decrease the size of the sample but remain reasonably small. Following these observations, we argue that a large sample size (as in [5]) 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

For a balanced dataset obtained with the method described in Section 3.1, Fig. 3 shows for the $\mathcal{Z}$ space of PGGAN CelebAHQ that a stronger regularization (smaller value of C) leads to smaller entanglement, while the effect on the target attribute remains almost unchanged. Similar results are reported in the supplementary material for the $\mathcal{Z}$ and $\mathcal{W}$ spaces of StyleGAN FFHQ. 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). We find that this direction gives the best performances.

4.4 Identity preservation

Table 4 shows a quantitative evaluation of identity preservation. We compare our results with InterFaceGAN for the different models. Some attributes affect the identity more than others, in particular "gender". For InterFaceGAN, we observe that the identity is less well-preserved when manipulating "age" and "glasses", which might be explained by the fact that they are entangled with "gender" (cf. Tables 1, 2). For "smile", which is a well-disentangled attribute, we perform slightly better or on par with InterFaceGAN for directions with similar effects (cf. Tables 1, 2, 3). This suggests that our directions preserve the identity well. Our results in the $\mathcal{W}$ space of StyleGAN for the attributes "glasses" and "age" are quite below those of InterFaceGAN, which is probably due to our directions having more effect (see *e.g.* Fig. 6 where we end up with quite occluding sunglasses).

	Glasses			Gender			Smile			Age		
	P	S $\mathcal{Z}$	S $\mathcal{W}$	P	S $\mathcal{Z}$	S $\mathcal{W}$	P	S $\mathcal{Z}$	S $\mathcal{W}$	P	S $\mathcal{Z}$	S $\mathcal{W}$
IfGAN	0.59 ± 0.14	0.72 ± 0.13	**0.70** ± 0.11	0.56 ± 0.14	0.66 ± 0.14	**0.71** ± 0.13	0.74 ± 0.10	0.86 ± 0.09	**0.81** ± 0.09	0.66 ± 0.15	0.70 ± 0.14	**0.76** ± 0.11
Ours	**0.77** ± 0.15	**0.73** ± 0.13	0.63 ± 0.12	**0.62** ± 0.14	**0.74** ± 0.13	0.68 ± 0.13	**0.79** ± 0.08	**0.87** ± 0.08	0.78 ± 0.08	**0.78** ± 0.10	**0.73** ± 0.12	0.69 ± 0.11

Table 4: Identity preservation (higher is better) for manipulation in $\mathcal{Z}$ of PGGAN CelebAHQ (P) and in $\mathcal{Z}$ and $\mathcal{W}$ of StyleGAN FFHQ (S $\mathcal{Z}$ resp. S $\mathcal{W}$).

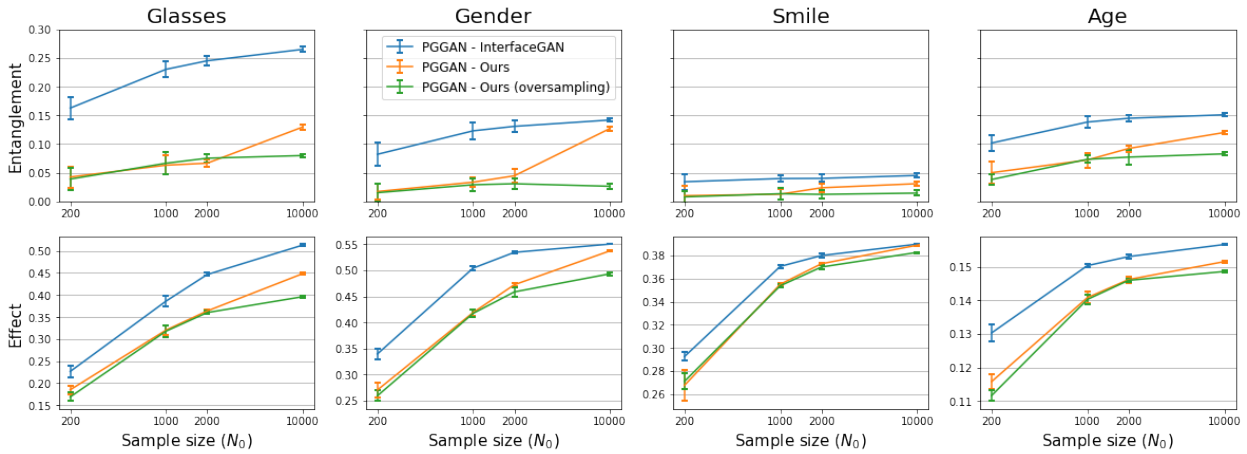


Figure 2: Influence of sample size on the overall entanglement (top) and desired effect (bottom) associated with a direction in the $\mathcal{Z}$ space of PGGAN CelebAHQ.

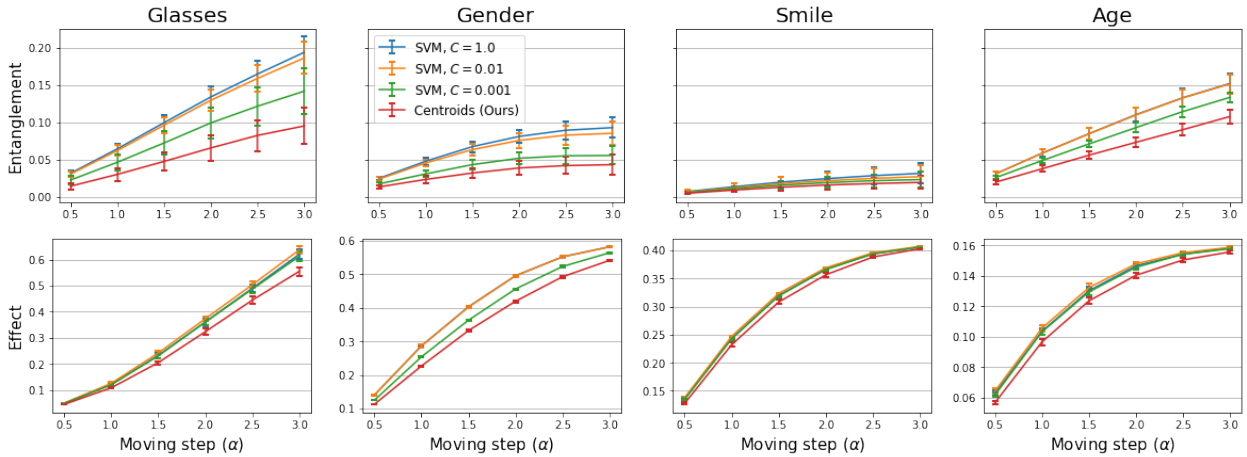


Figure 3: Influence of SVM regularization on the overall entanglement (top) and desired effect (bottom) associated with a semantic direction in the $\mathcal{Z}$ space of PGGAN CelebAHQ.

4.5 Qualitative results

In Fig. 4, we show qualitative results in the $\mathcal{Z}$ space of PGGAN CelebAHQ and StyleGAN FFHQ. We find that the images obtained with InterFaceGAN show significant entanglement. On par with the quantitative results, we notice that the direction "glasses" is entangled with "age" and "gender", the direction "gender" also affects the attributes "age" and "glasses" and the direction "age" tends to feminize. In contrast, our directions better preserve the non-target attributes. In Fig. 5 and Fig. 6 we show qualitative results for different amplitude values. Additional results for the different attributes are presented in the supplementary material.

5 Discussion

While our method balances the sample to decorrelate the attributes, we observed that the resulting directions in latent space are quasi-orthogonal (see supplementary material), 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 [2] 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 [5] 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

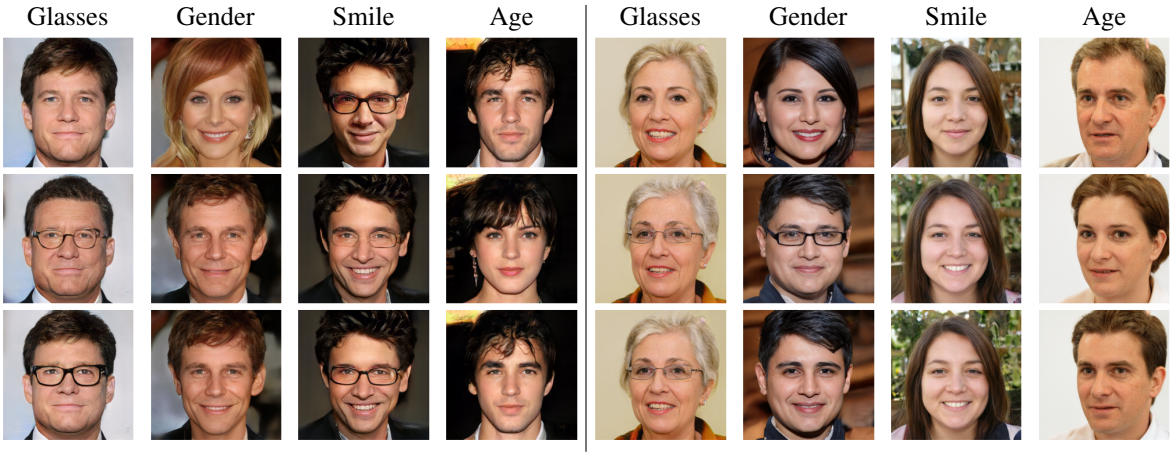


Figure 4: Qualitative results in $\mathcal{Z}$ space of PGGAN CelebAHQ (left) and StyleGAN FFHQ (right). First row: input image. Second row: InterFaceGAN. Third row: our method.

Figure 5: Continuous manipulation for the attribute "age" in $\mathcal{Z}$ space of PGGAN CelebAHQ.

Figure 6: Continuous manipulation for the attribute "glasses" in $\mathcal{W}$ space of StyleGAN FFHQ.

approach can prove beneficial to other works on GAN control that rely on sampling in the latent space. Two 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 continuous attributes that are binarized (*e.g.* "age") and might not be a problem in practice.

6 Conclusion

We focused on the identification of linear 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.

References

[1] 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*, volume 27. Curran Associates, Inc., 2014.

[2] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. GANSpace: Discovering interpretable gan controls. In *Advances in Neural Information Processing Systems*, 2020.

[3] Antoine Plumerault, Hervé Le Borgne, and Céline Hudelot. Controlling generative models with continuous factors of variations. In *International Conference on Learning Representations*, 2020.

[4] Ali Jahanian, Lucy Chai, and Phillip Isola. On the "steerability" of generative adversarial networks. In *International Conference on Learning Representations*, 2020.

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

[6] Peiye Zhuang, Oluwasanmi O Koyejo, and Alex Schwing. Enjoy your editing: Controllable GANs for image editing via latent space navigation. In *International Conference on Learning Representations*, 2021.

[7] Ceyuan Yang, Yujun Shen, and Bolei Zhou. Semantic hierarchy emerges in deep generative representations for scene synthesis. *International Journal of Computer Vision*, 2020.

[8] 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 (CVPR)*, pages 1532–1540, June 2021.

[9] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the GAN latent space. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9786–9796. PMLR, 13–18 Jul 2020.

[10] Ben Hutchinson, Emily Denton, Margaret Mitchell, and Timnit Gebru. Detecting bias with generative counterfactual face attribute augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019.

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

[12] 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 (CVPR)*, June 2019.

[13] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[14] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations*, 2016.

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

[16] 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 (ICCV)*, October 2019.

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

[18] Rameen Abdal, Peihao Zhu, Niloy J. Mitra, and Peter Wonka. StyleFlow: Attribute-conditioned exploration of StyleGAN-generated images using conditional continuous normalizing flows. *ACM Transaction on Graphics*, 40(3), May 2021.

[19] Xianxu Hou, Xiaokang Zhang, Linlin Shen, Zhihui Lai, and Jun Wan. GuidedStyle: Attribute knowledge guided style manipulation for semantic face editing. *ArXiv*, abs/2012.11856, 2020.

[20] Hui-Po Wang, Ning Yu, and Mario Fritz. Hijack-GAN: Unintended-use of pretrained, black-box GANs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7872–7881, June 2021.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[22] Zongze Wu, Dani Lischinski, and Eli Shechtman. StyleSpace analysis: Disentangled controls for StyleGAN image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12863–12872, June 2021.

[23] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 67–74, 2018.