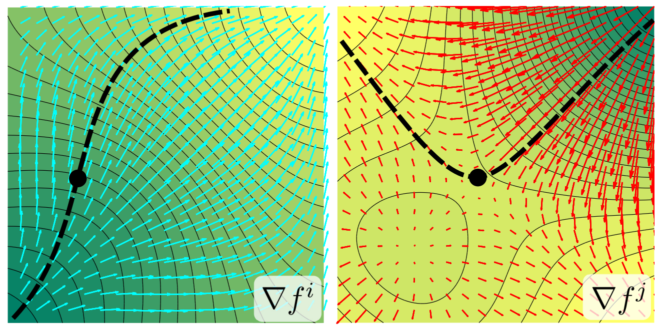
WarpedGANSpace: Finding non-linear RBF paths in GAN latent space

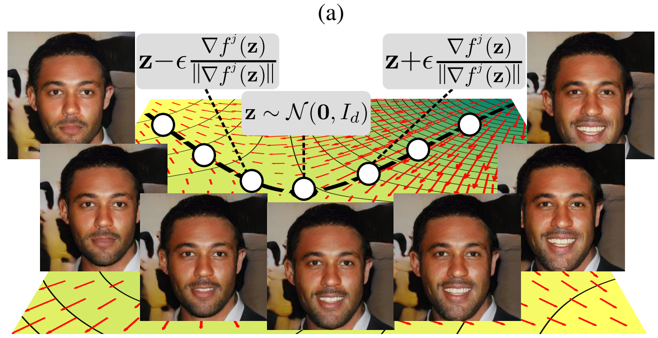
Christos Tzelepis, Georgios Tzimiropoulos, Ioannis Patras
Queen Mary University of London
Mile End road, E1 4NS London, UK
{c.tzelepis, g.tzimiropoulos, i.patras}@qmul.ac.uk

Abstract

This work addresses the problem of discovering, in an unsupervised manner, interpretable paths in the latent space of pretrained GANs, so as to provide an intuitive and easy way of controlling the underlying generative factors. In doing so, it addresses some of the limitations of the state-of-the-art works, namely, a) that they discover directions that are independent of the latent code, i.e., paths that are linear, and b) that their evaluation relies either on visual inspection or on laborious human labeling. More specifically, we propose to learn non-linear warpings on the latent space, each one parametrized by a set of RBF-based latent space warping functions, and where each warping gives rise to a family of non-linear paths via the gradient of the function. Building on the work of [34], that discovers linear paths, we optimize the trainable parameters of the set of RBFs, so as that images that are generated by codes along different paths, are easily distinguishable by a discriminator network. This leads to easily distinguishable image transformations, such as pose and facial expressions in facial images. We show that linear paths can be derived as a special case of our method, and show experimentally that non-linear paths in the latent space lead to steeper, more disentangled and interpretable changes in the image space than in state-of-the art methods, both qualitatively and quantitatively. We make the code and the pretrained models publicly available at: `https://github.com/chi0tzp/WarpedGANSpace`.



(a)



(b)

Figure 1: (a) Warpings of vector space $\mathbb{R}^d$ due to two RBF functions, f^i and f^j, lead to different non-linear paths in $\mathbb{R}^d$ for any given $\mathbf{z}\in\mathbb{R}^d$ (dashed bold lines) via their gradients, ∇f^i and ∇f^j. Solid black lines represent isohypses of the warpings and the colored vectors represent the vector fields induced by their gradients. (b) Illustration of a non-linear path due to warping f^j, starting from a latent code $\mathbf{z}$ and moving along the gradient ∇f^j by steps of magnitude ϵ.

1. Introduction

Generative Adversarial Networks (GANs) [10] have emerged as the leading generative learning paradigm, exhibiting clear superiority in terms of the quality of generated realistic and aesthetically pleasing images [25, 3, 17, 18, 19]. However, despite their generative efficiency, GANs do not provide an inherent way of comprehending or controlling the underlying generative factors. To address this, the research community has directed its efforts towards studying the structure of GAN's latent space [28, 30, 18, 33, 7, 37, 1, 34, 35, 9, 14, 32, 26, 11]. These works study the structure of GAN's latent space and attempt to find interpretable directions on it; that is, directions sampling across which are expected to generate images where only a few (ideally one) factors of variations are "activated". Meaningful human-interpretable directions can refer to either domain-specific factors (e.g., facial expressions [28]) or domain-agnostic factors (e.g., zoom scale [14, 26, 32]).

Several methods adopt a supervised learning framework, and discover directions in the latent space that align well to factors controlled by supervision. In this line of research, [31, 9, 18] supervision is in the form of labels assigned to the generated images, either by explicit human annotation, or by the use of pretrained semantic classifiers. Recent works, such as [32, 14, 26], steer the directions in the latent space so as to align well with controllable manipulations in the image space (e.g., zoom). Those works are limited by the fact that the factors are assumed to be known and by practical issues in generating the supervisory signals.

Another line of research imposes unsupervised constraints in the directions in the latent space. GANSpace [11] performs PCA on deep features at the early layers of the generator and finds directions in the latent space that best map to those deep PCA vectors, arriving at a set of non-orthogonal directions in the latent space. Similarly to other methods, this is a very demanding training process that requires drawing large numbers of random latent codes and regressing the latent directions. Similarly, Voynov and Babenko [34] proposed an unsupervised method to discover linear interpretable latent space directions. While the unsupervised learning framework has interest, current works make the hard assumption that the discovered directions are isotropic in the latent space, leading to linear paths. Furthermore, despite the fact that these works lead to more complex directions, compared to methods that do not use any optimization at all (e.g., [14, 32]), the evaluation of the obtained results are either left to subjective visual inspection (e.g., [11]) or relies on laborious human labeling [34].

In this work, we propose to learn non-linear warping functions on the latent space, each one parametrized by a set of RBF-based latent space warping operations, and where each warping function f^k gives rise to a family of non-linear paths via its gradient. More precisely, at each latent code $\mathbf{z} \in \mathbb{R}^d$, the gradient of the warping function $\nabla f^k(\mathbf{z})$ gives the direction along the k-th family of paths – clearly, the gradient of f^k is not isotropic in $\mathbb{R}^d$, giving rise to non-linear paths. An example is shown in Fig. 1, where two RBF-warping functions, f^i and f^j, are depicted together with two distinct non-linear paths. Building on the work of [34], that discovers linear paths, we optimize the trainable parameters of the RBFs, so as that images that are generated by codes along paths of different families, f^k, are easily distinguishable by a discriminator network (Fig. 2) – this leads to easily distinguishable image transformations, such as pose and facial expressions in facial images (Fig. 1b). We show that [34], which learns linear paths, can be derived as a special case of our method and we perform extensive comparisons with state-of-the art methods both qualitatively and quantitatively.

For a quantitative evaluation, we propose to utilize trained classifiers that assign attributes to the generated images and propose a framework that monitors the correlation between paths in the latent space, and the corresponding changes/paths in the attribute space so as to determine how correlated are paths along certain warping functions to certain attributes. We experimentally show that the proposed non-linear paths in the latent space lead to more disentangled and more interpretable changes in the image space than in state-of-the art methods. In addition, we show that for paths of the same length in the latent space, our method is able to produce much larger changes in the attribute space in comparison to the linear one, i.e., the generated attribute paths are much more steep, and that we are able to generate larger attribute changes before the quality of the generated images deteriorates.

The main contributions of this paper can be summarized as follows:

- We propose an unsupervised and model-agnostic method for discovering non-linear interpretable paths on the latent space of pretrained GANs by using RBF-based warping functions. We derive the case of linear paths as a special case and learn a set of such warping functions so that the corresponding image transformations are distinguishable to each other.
- We propose a quantitative evaluation protocol for measuring the interpretability/disentanglement of paths in the latent space, by analysing the corresponding changes to attributes in the generated images, as those are measured by pretrained semantic classifiers (e.g., pretrained face attribute networks).
- We apply our method to four pretrained GANs (i.e., SN-GAN [25], BigGAN [3], ProgGAN [17], and StyleGAN2 [19]) and compare our non-linear paths to linear ones [34, 11], both qualitatively and quantitatively. We show that in comparison to state-of-the-art, our method produces steeper, more disentangled, and longer paths in the attribute space.

2. Related Work

Disentanglement in generative learning A disentangled representation in the context of generative learning can be defined as one where single latent units are sensitive to changes in single generative factors, while being relatively invariant to changes in other factors [2]. Imposing disentanglement in the latent space of a generative learning method has drawn significant attention by the research community in recent years. These works typically refer to the notion of a disentangled latent space [4, 13, 23, 22, 29, 36], in the context of either VAE (e.g., [36, 13]) or GAN (e.g., [4, 23]) and they typically try to improve the architectures and the

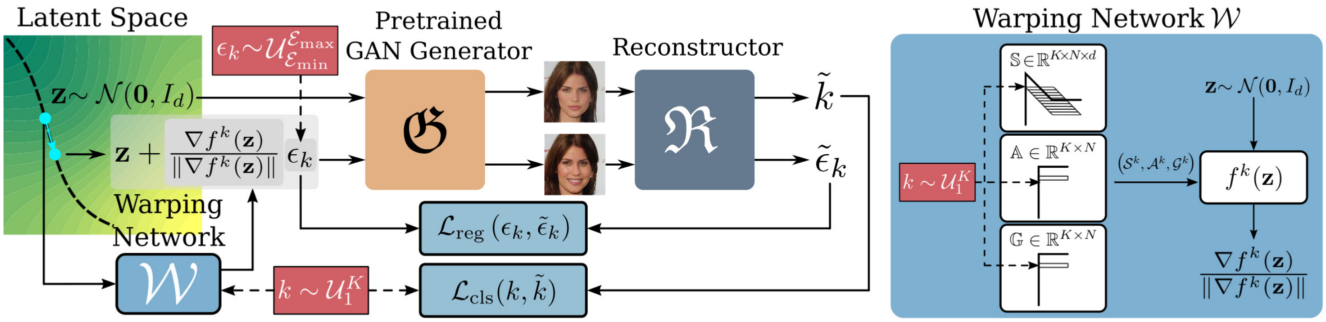


Figure 2: Overview the proposed method: A latent code $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, I_d)$ is shifted by a vector induced by a warping function f^k implemented by the warping network $\mathcal{W}$ after choosing the corresponding support set $\mathcal{S}^k$, weights $\mathcal{A}^k$, and parameters $\mathcal{G}^k$. The pair of latent codes, $\mathbf{z}$ and $\mathbf{z} + \epsilon_k \frac{\nabla f^k(\mathbf{z})}{\|\nabla f^k(\mathbf{z})\|}$, are then fed into the generator $\mathfrak{G}$ in order to produce two images. The reconstructor $\mathfrak{R}$ is optimized to reproduce the signed shift magnitude ϵ_k and predict the index k of the support set used.

training protocols of standard generative methods in order to obtain latent spaces where generative factors are disentangled. While these works provide comprehensive theoretical insights, they are typically applied to toy or low-resolution datasets and exhibit inferior results in terms of generation quality and diversity compared to state-of-the-art GANs, such as ProgGAN [17] or StyleGAN2 [19].

Discovering interpretable paths in pretrained GAN generators Since the early days of GANs, it has been shown that the GAN latent space often exhibits semantically meaningful vector space arithmetic. Radford et al. [27] showed that there exist latent directions corresponding to adding smiles or glasses on faces. This paved the way for the development of methods that would facilitate image editing and has since received significant research attention. Some works [9, 30, 18] require explicit human-provided supervision to identify interpretable directions in the latent space. More specifically, [30, 18] use classifiers, pretrained on the CelebA dataset [24], in order to predict certain face attributes. These classifiers are then used to produce pseudo-labels for a large number of generated images and their latent codes. Based on these pseudo-labels, a separating hyperplane is learned in the latent space giving rise to a direction that captures the corresponding attribute. Plumerault et al. [26] also solve an optimization problem in the latent space for maximizing the score of the pretrained model to predict image memorability and then find the directions that increases memorability. By contract to the above works, our method is trained in an unsupervised manner.

Some recent works [14, 26, 32] seek those vectors in the latent space that correspond to controlled image augmentations such as zoom or translation. While these approaches have interest, they can find only the directions capturing the transformations that they have been trained on. By contrast, our method can discover non-linear paths that correspond to more complex generative factors (e.g., skin color, age, etc.).

Finally, our method is closely related to those of [34, 11], since we are also learning a set of interpretable paths in an unsupervised and model-agnostic manner. More specifically, Voynov and Babenko [34] optimize a set of linear interpretable directions, modeled by a set of vectors in the latent space, and they evaluate the performance of their method using the judgements of eleven human assessors. GANSpace [11] is trained in an unsupervised manner in order to discover meaningful directions by using PCA on deep features of the generator. This method seeks linear directions in the latent space that best map to those deep PCA vectors, and results in a set of non-orthogonal directions. Similarly to other methods discussed above, it also requires a very demanding training procedure (drawing random latent codes and regressing the latent directions), while they provide only qualitative evaluation results.

In contrast to these works, our method discovers non-linear paths in the latent space of a pretrained GAN generator in an unsupervised manner. Moreover, in order to lift the obvious limitations introduced by manual labeling of the discovered paths, we propose a quantitative and automatic evaluation protocol that obtains the most interpretable paths in terms of correlation with a certain number of attributes.

3. Proposed Method

In this section, we present our method for discovering K non-linear interpretable paths on the latent space of a pretrained GAN generator, by learning K warping functions, $f^1, \ldots, f^K$, the gradients of which define the directions of the paths at each latent code $\mathbf{z} \in \mathbb{R}^d$. More specifically, we transform $\mathbb{R}^d$ by $f^k \colon \mathbb{R}^d \to \mathbb{R}$ that is parameterized as a weighted sum of RBFs, and for any given $\mathbf{z} \in \mathbb{R}^d$ we move along the path belonging to the k-th family of paths by following the direction of $\nabla f^k(\mathbf{z})$. In order to obtain interpretable paths, we adopt the framework of [34] and learn warping functions that give families of paths that lead to image transformations that are distinguishable to each other by a discriminator/reconstructor. The parameters of the warping function and of the reconstructor/discriminator network are optimized jointly. By contrast to [34] and other methods in the literature, the warping functions may lead to non-linear paths, and the linear ones can be obtained for specific values of the parameters. An overview of the proposed method is given in Fig. 2.

3.1. Vector space warping and traversal

Given a vector space $\mathbb{R}^d$, we define $f\colon \mathbb{R}^d \to \mathbb{R}$ as a weighted sum of parametric Gaussian RBFs given by

$$f(\mathbf{z}) = \sum_{i=1}^{N} \alpha_i \exp\left(-\gamma_i \|\mathbf{z} - \mathbf{s}_i\|^2\right), \tag{1}$$

where $\alpha_i \in \mathbb{R}$, $\gamma_i \in \mathbb{R}_+$, and $\mathbf{s}_i \in \mathbb{R}^d$, denote the weight, the scale, and the center of the i-th RBF, respectively. Geometrically, f transforms each point $\mathbf{z}$ of the given vector space $\mathbb{R}^d$ into a $(d+1)$-dimensional point $(\mathbf{z}, f(\mathbf{z}))$ that lies on a d-dimensional manifold. We define this transformation as a *warping* of the vector space $\mathbb{R}^d$. Also, hereby, we will be referring to the centers of the RBFs as the *support vectors*, driven by the geometric intuition that they "support" the induced warping of the space, and we will be using the term *support set* to refer to the set of support vectors, $\mathcal{S} = \{\mathbf{s}_i \in \mathbb{R}^d, i = 1, \ldots, N\}$. The corresponding weights and γ parameters will be hereby referred to as the sets $\mathcal{A} = \{\alpha_i \in \mathbb{R}, i = 1, \ldots, N\}$ and $\mathcal{G} = \{\gamma_i \in \mathbb{R}_+, i = 1, \ldots, N\}$, respectively. Then, different support sets will in general lead to different warpings of a given vector space.

The above warping operation is differentiable and its gradient is given analytically as follows

$$\nabla f(\mathbf{z}) = -2 \sum_{i=1}^{N} \alpha_i \gamma_i \exp\left(-\gamma_i \|\mathbf{z} - \mathbf{s}_i\|^2\right) (\mathbf{z} - \mathbf{s}_i). \tag{2}$$

Thus, given an arbitrary $\mathbf{z}$, $\nabla f(\mathbf{z})$ defines a (local) direction, which we use in order to define a curve in $\mathbb{R}^d$. More specifically, for any $\mathbf{z} \in \mathbb{R}^d$ and sufficiently small shift magnitude ϵ, we define a continuous curve in $\mathbb{R}^d$ induced by the warping operation f using (2) by shifting $\mathbf{z}$ by

$$\delta\mathbf{z} = \epsilon \frac{\nabla f(\mathbf{z})}{\|\nabla f(\mathbf{z})\|}. \tag{3}$$

In Fig. 1a, we illustrate this for a given vector space $\mathbb{R}^d$ and two warpings, f^i and f^j, which lead to two different non-linear paths in $\mathbb{R}^d$ for any given $\mathbf{z}$ (dashed bold lines). In this figure, thin solid lines represent level sets of the warpings, while the vector fields represent their gradients.

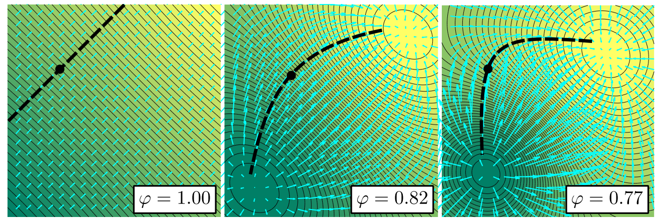


Figure 3: Illustration of gradient fields for warpings of bipolar RBFs, for small, medium, and large γ parameter values.

3.2. Learning non-linear interpretable curves in GAN's latent space

Following the discussion above, given a pretrained GAN's latent space, which is typically modeled as a d-dimensional vector space $\mathcal{Z} \subseteq \mathbb{R}^d$, we may model a set of different warpings by a set of support sets $\{\mathcal{S}^k\}$, along with the corresponding weights $\{\mathcal{A}^k\}$ and γ parameters $\{\mathcal{G}^k\}$, $k = 1, \ldots, K$. We embed the support sets into the *support tensor* $\mathbb{S} \in \mathbb{R}^{K \times N \times d}$, and the weights and γ parameters into the matrices $\mathbb{A} \in \mathbb{R}^{K \times N}$ and $\mathbb{G} \in \mathbb{R}^{K \times N}$, respectively. Then, each support set, along with the corresponding weights and γ parameters, leads to a specific warping of the latent space via the function f^k defined by (1), whose gradient is given analytically by (2). Thus, for each $(\mathcal{S}^k, \mathcal{A}^k, \mathcal{G}^k)$, $k = 1, \ldots, K$, we define a vector field on the latent space, which we use to traverse it using (3).

Here, we define each warping to be given by a set of pairs of "bipolar" support vectors, i.e., pairs, that have opposite weights α and equal scale γ. In this formulation, γ controls the degree of non-linearity of the path, where very small γ lead to linear paths, similar to [34]. This is illustrated in Fig. 3, where the vector fields for two bipolar support vectors with different values of γ are depicted.

Finally, let us note that in contrast to the global linear directions discovered by [34, 11], in our case the directions along each warping are different for different latent codes. That is, as shown in (3), the gradient and the shift vector depend on the latent code itself. This anisotropic behaviour of the proposed method reflects our intuition that interpretable paths do not necessarily have the same direction at every region of the latent space.

Linear directions as a special case In this section we will show that the method of [34] can be derived as a special case of our method. We first note that the framework of [34] that discovers linear directions encoded in the columns of a matrix A can be derived in the special case that the warping functions are linear in $\mathbf{z}$, that is, $f(\mathbf{z}) = A^\top \mathbf{z}$. In that case, the direction along the k-th direction is given by $\delta\mathbf{z} = \nabla f^k(\mathbf{z}) = \mathbf{a}_k$, where $\mathbf{a}_k$ is the k-th column of A.

It is straightforward to show that this solution can be ob-

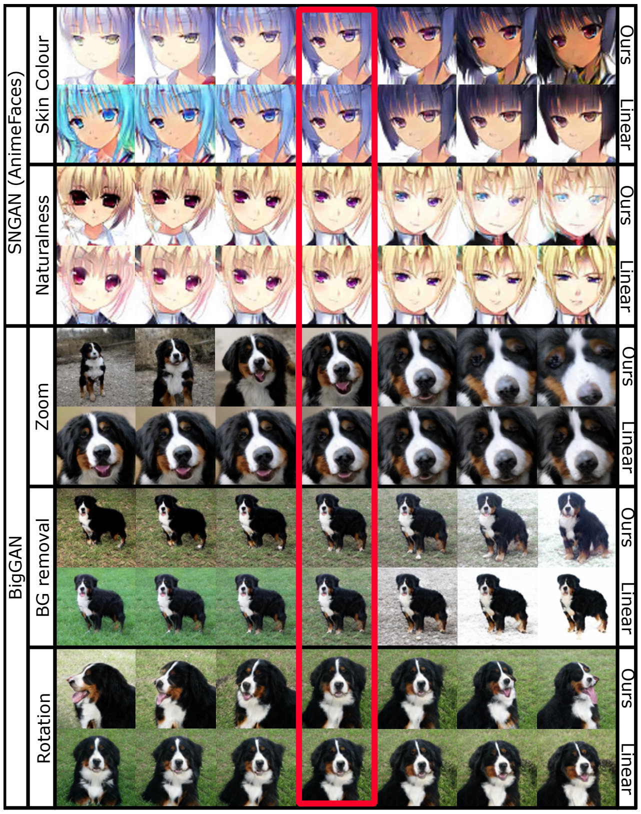


Figure 4: Interpretable paths discovered by our method (non-linear) compared to the corresponding linear ones discovered by [34] for SN-GAN and BigGAN.

tained in our formulation, when each of the RBF-warping is given by pairs of bipolar RBFs, i.e, pairs of support vectors with opposite α and the same γ, when the value of γ is sufficiently small. In what follows we give the proof for the simple case of a single bipolar pair, in the special case that $s_1 = -s_2 = s$. In that case, (2) can be written as $\nabla f^k(\mathbf{z}) = -2\alpha^k\gamma^k \exp\left(-\gamma^k\|\mathbf{z}-\mathbf{s}^k\|^2\right)\left(\mathbf{z}-\mathbf{s}^k\right) + 2\alpha^k\gamma^k \exp\left(-\gamma^k\|\mathbf{z}+\mathbf{s}^k\|^2\right)\left(\mathbf{z}+\mathbf{s}^k\right)$, which, for sufficiently small γ^k, leads to $\nabla f^k(\mathbf{z}) = 4\alpha^k\gamma^k\mathbf{s}^k$. Then, the shift in the latent space, given by (3), is written as

$$\delta\mathbf{z} = \epsilon\frac{4\alpha^k\gamma^k\mathbf{s}^k}{\|4\alpha^k\gamma^k\mathbf{s}^k\|} = \epsilon\frac{\mathbf{s}^k}{\|\mathbf{s}^k\|}. \tag{4}$$

In this case, the derivative of the k-th warping function at $\mathbf{z}$ is independent of $\mathbf{z}$ and equal to a constant vector.

It is straightforward to show that linear directions can be obtained also in the more general case that each of the warping functions is given by several bipolar support vectors, each with a small γ. It is also the case that such parameters could be found by the optimization process, if they lead to discernible image transformations.

3.3. Learning process

An overview of the learning process is presented in Fig. 2. We use a pretrained generator $\mathfrak{G}$ and learn a) the parameters of a warping network $\mathcal{W}$ that generates paths in the latent space of $\mathfrak{G}$, and b) the parameters of a reconstructor network $\mathfrak{R}$ that recognises the index k of the warping that generated the changes between a pair of images. The trainable modules of our method are the following:

Warping Network The warping network $\mathcal{W}$ is parametrized by a set of triplets, $(\mathcal{S}^k, \mathcal{A}^k, \mathcal{G}^k)$, of the support set $\mathcal{S}^k$, and the corresponding weights $\mathcal{A}^k$ and γ parameters $\mathcal{G}^k$, $k = 1, \ldots, K$. Each such triplet gives rise to a warping of the latent space $\mathbb{R}^d$, and thus, to a non-linear path for any given latent code $\mathbf{z} \in \mathbb{R}^d$. $\mathcal{W}$ is implemented by standard layers and is differentiable.

Reconstructor A reconstructor $\mathfrak{R}$ is a model that we use in order to distinguish the image transformations that are induced by the different support sets (i.e., the different latent space warpings). As shown in Fig. 2, the input to the reconstructor is a pair of images, $\mathfrak{G}(\mathbf{z})$ and $\mathfrak{G}\left(\mathbf{z} + \epsilon_k\frac{\nabla f^k(\mathbf{z})}{\|\nabla f^k(\mathbf{z})\|}\right)$. The reconstructor's goals are i) to predict which support set gave rise to the transformation at hand, i.e., recognise the index k and ii) to reproduce the magnitude of the shift in the latent space; that is, predict ϵ_k. In the experiments, we use the LeNet [21] backbone for SN-GAN (MNIST and AnimeFaces datasets) and ResNet-18 [12] for BigGAN (ImageNet), ProgGAN (CelebA-HQ), and StyleGAN2 (FFHQ). We modify the input channels of the reconstructor so as it receives pairs of images (i.e., we concatenate the input image pair along channels dimension). Finally, we define two output "heads", one for predicting the index (classification), and the other for predicting the shift magnitude (regression).

Optimization objective The optimization problem that we solve is as follows

$$\min_{\mathbb{S},\mathbb{A},\mathbb{G},\mathfrak{R}} \mathbb{E}_{\mathbf{z},k,\epsilon}\left[\mathcal{L}_{\text{cls}}(k,\tilde{k}) + \lambda\mathcal{L}_{\text{reg}}(\epsilon,\tilde{\epsilon})\right], \tag{5}$$

where $\mathcal{L}_{\text{cls}}$ denotes the classification loss term where we use the cross-entropy function, $\mathcal{L}_{\text{reg}}$ denotes the regression loss terms where we use the mean absolute error, and λ is a weighting coefficient. We note that the objective function is differentiable with respect to the support vectors, weights α and γ parameters, allowing us to learn not only the positions of the support vectors, but also their weights, and/or γ parameters. To ensure the positivity of γ we learn its logarithm. As discussed above, for each warping we learn a set of bipolar pairs of support vectors.

During training, we generate pairs of images $\mathfrak{G}(\mathbf{z})$ and $\mathfrak{G}\left(\mathbf{z} + \epsilon_k\frac{\nabla f^k(\mathbf{z})}{\|\nabla f^k(\mathbf{z})\|}\right)$, where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, I_d)$, k is a warping

Figure 5: Non-linear interpretable paths automatically discovered by our method in StyleGAN2's [19] $\mathcal{W}$-space.

Method	GAN				
	SNGAN (MNIST)	SNGAN (Anime)	BigGAN	ProgGAN	StyleGAN2
Random	46.0	85.0	76.0	60.0	-
Coord	48.0	89.0	66.0	82.0	-
Linear [34]	88.0	99.0	85.0	90.0	-
Ours	**98.4**	**99.8**	**92.6**	**99.3**	**99.8**

Table 1: Reconstructor accuracy (%) of the proposed method compared to [34] (linear directions), random latent direction and latent directions aligned with axes, for various GAN generators pretrained on the given datasets.

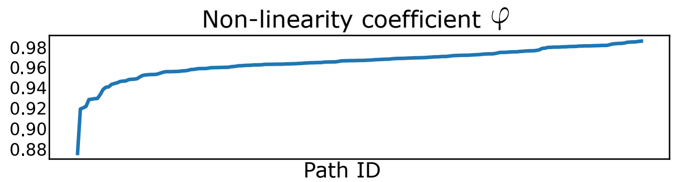


Figure 6: Illustration of the non-linearity coefficient φ for the discovered paths obtained by our method for ProgGAN.

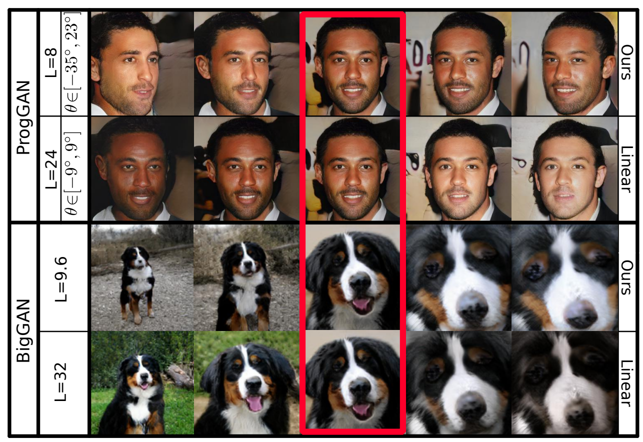


Figure 7: Effect of traversal path length L in comparison with the linear case [34].

function index uniformly sampled in $\{1, \ldots, K\}$, and ϵ_k is a scalar sampled uniformly in $\mathcal{U}_{\mathcal{E}_{\min}}^{\mathcal{E}_{\max}} = \mathcal{U}[-\mathcal{E}_{\max}, -\mathcal{E}_{\min}] \cup \mathcal{U}[\mathcal{E}_{\min}, \mathcal{E}_{\max}]$. The pair of images is fed to the reconstructor where the loss is calculated and the gradients are back-propagated to the warping network and the reconstructor.

4. Experiments

Overview of results In this section we will present the experimental evaluation of the proposed method and provide qualitative and quantitative comparisons with state-of-the-art methods. We will first show that in comparison to [34] our method finds paths in the latent space that produce changes in the image space that are easier to be distinguished by a discriminating network – this is achieved consistently across several GANs that are pretrained on different datasets (Table 1). We will then show that in comparison to the state-of-the-art, our method finds paths in the latent space, that produce more distinguishable, more disentangled and larger changes in the generated images. We will first show that qualitatively by presenting images generated along paths of equal length in the latent space for different methods (Fig. 4,8,5) and observe the generated variations they produce in the image space. We will subsequently show this quantitatively (Table 2), by estimating semantic attributes (e.g., rotations, smile, etc.) in the generated images, and report the correlations and ranges as we follow different paths in the latent space. Finally, we will show that our method finds paths on the latent space that correspond to steeper changes/paths in the attribute space, and therefore allows for better, controllable generation without arriving at latent space regions of low density and, thus, at quality degradation or distortions (Fig. 7,8).

Pretrained GAN generators and datasets We evaluate the proposed method using the following pretrained GANs: a) Spectrally Normalized GAN (SN-GAN) [25] trained on MNIST [20] and AnimeFaces [15], b) BigGAN [3] trained on ImageNet [5], c) ProgGAN [17] trained on CelebA-HQ [24], and d) StyleGAN2 [19] trained on FFHQ [19].

Paths with more distinguishable changes in the image space We first show that a reconstructor that discriminates images according to the warping in the latent space that generated them, i.e., estimates the index of the warping function, has better classification performance than in the corresponding linear case [34]. This is an indication that the paths that are generated by our method can be discriminated more effectively and therefore are more likely to be

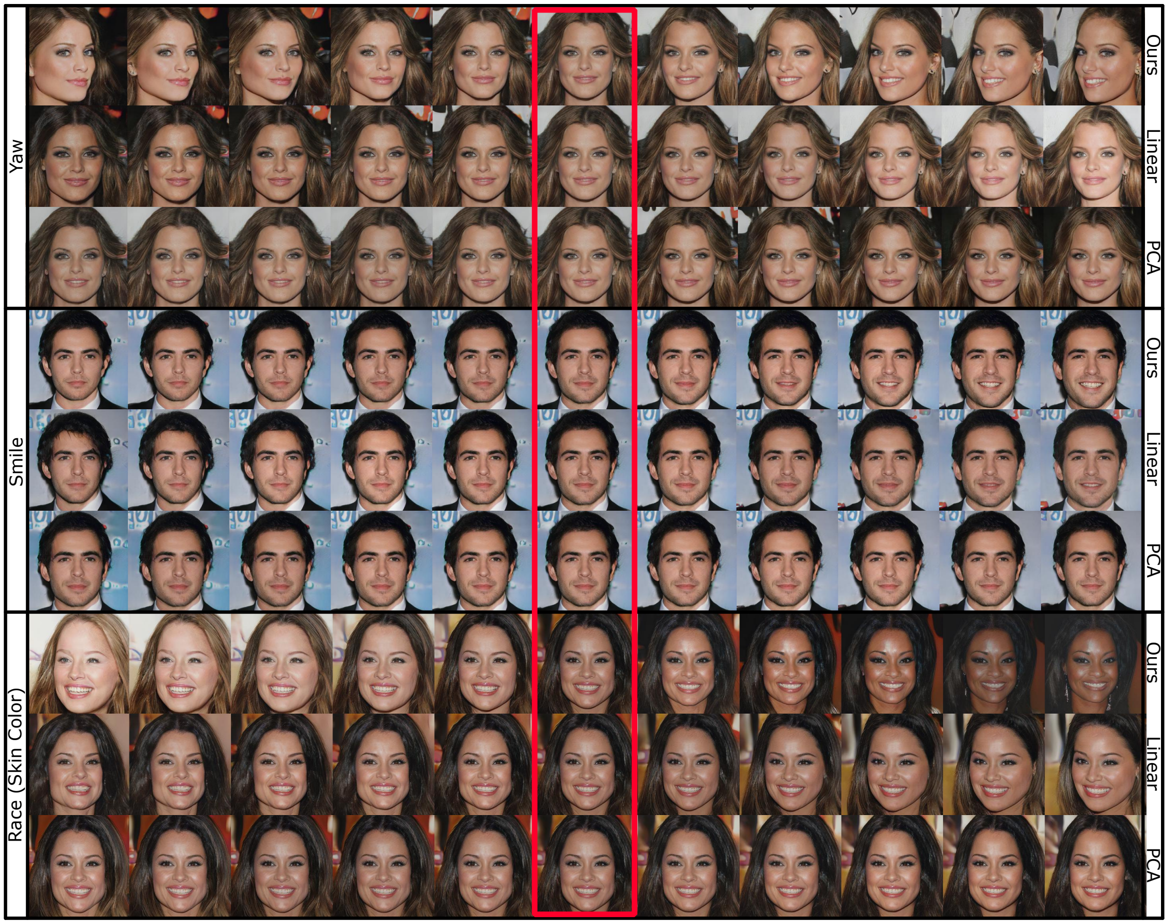


Figure 8: Automatically discovered non-linear (ours – first row) and linear (Voynov and Babenko [34] – second row, GANSpace [11] – third row) interpretable paths in ProgGAN's [17] latent space.

more interpretable. The results are summarised in Table 1 and are consistent across several pretrained GANs.

Interpretable paths with steeper and more disentangled changes in the image space – qualitative evaluation We then show qualitatively that the proposed method finds interpretable paths in the latent space that are similar to the ones reported in [34], but exhibit larger variations in the captured generative factors. More specifically, for a given method that discovers a set of paths, that is, linear in the cases of [34, 11] or non-linear in our case, in the latent space of a pretrained GAN, we generate an image sequence for each path, starting from a random latent code and "walking" towards the positive and the negative ways of the path for a certain amount of steps. This gives rise to an image sequence that shows how the learned path at hand affects the generation. For fair comparison, the step size and therefore the path length, is the same for all methods.

In Fig. 4 we show the generated images along manually selected directions found by our method and the method of [34] on SN-GAN (AnimeFaces). In the same figure, we show three interpretable paths discovered by our method, namely zoom, background removal, and rotation, in comparison with the corresponding ones reported in [34] – we note that these are the directions chosen in [34] and that we generate the paths using the publicly available models provided by the authors. We can clearly see that in both cases, the paths found by our method produce larger changes in the image space and larger variations in the content.

In Fig. 8 we show paths discovered on the latent space of ProgGAN [17], that is trained on CelebA-HQ [24]. For this method we report the directions that are most correlated with three attributes, namely *yaw*, *smile*, and *race*, with the correlations estimated with a method we will describe below. We compare with the corresponding linear directions obtained by [34, 11] and we note that our method both leads to greater variation in the respective generative factors (e.g., larger rotation angles) for the same traversal lengths in the latent space, but also that we are able to produce more disentangled generations. This is apparent in Fig. 8, where,

Table 2: Comparison of the proposed method (non-linear latent paths) to [34] (linear latent directions) and GANSpace [11] (linear PCA-based latent directions) in terms of $\mathcal{L}_1$-normalized correlation and range (r).

	ID	Yaw	Pitch	Smile	Race	Hair	r
Yaw	0.52	**0.32**	0.05	0.01	0.07	0.03	**43.66°**
Pitch	0.41	0.04	**0.38**	0.13	0.03	0.01	**22.53°**
Smile	0.24	0.03	0.07	**0.61**	0.03	0.03	**0.37**
Race	0.32	0.03	0.12	0.08	**0.29**	0.17	0.06
Hair	0.23	0.02	0.11	0.13	0.02	**0.49**	**0.28**

(a) Non-linear paths (Ours).

	ID	Yaw	Pitch	Smile	Race	Hair	r
Yaw	0.51	**0.24**	0.21	0.01	0.02	0.01	18.93°
Pitch	0.47	0.01	**0.25**	0.04	0.00	0.22	8.27°
Smile	0.24	0.01	0.04	**0.57**	0.05	0.09	0.28
Race	0.52	0.05	0.02	0.10	**0.31**	0.01	**0.16**
Hair	0.43	0.00	0.10	0.06	0.04	**0.36**	0.27

(b) Linear directions (Voynov and Babenko [34]).

	ID	Yaw	Pitch	Smile	Race	Hair	r
Yaw	0.47	**0.27**	0.04	0.13	0.03	0.06	17.65°
Pitch	0.45	0.05	**0.38**	0.09	0.02	0.01	7.48°
Smile	0.21	0.00	0.07	**0.55**	0.08	0.08	0.21
Race	0.35	0.11	0.02	0.12	**0.27**	0.12	0.10
Hair	0.44	0.05	0.06	0.03	0.08	**0.34**	0.15

(c) Linear PCA directions (GANSpace [11]).

for instance, changing *smile* attribute using our method preserves other generative factors better than [34, 11].

As noted, the length of the paths in the latent space is the same for all sequences and methods. To obtain a measure of the non-linearity of the generated paths, we calculate the ratio φ between the length of a path and the distance between its endpoints, and report the averages for all the traversals on a given warping. Clearly, for linear paths, $\varphi = 1$. The results are summarized in Fig. 6, where we plot (sorted) the values of φ for the discovered non-linear warpings for ProgGAN. An illustration is given in Fig. 3.

Non linear interpretable paths with steeper and more disentangled changes in the image space – quantitative evaluation In this section we will present our quantitative evaluation scheme, which we use for assessing the performance of our method and compare it to state-of-the-art [34, 11], for ProgGAN and StyleGAN2.

As discussed before, for a given method that discovers a set of interpretable paths; that is, linear in the cases of [34, 11] or non-linear in the case of the proposed method, in the latent space of a pretrained GAN generator, we generate an image sequence for each path, starting from a random latent code and "walking" towards the positive and the negative ways of the path for a certain amount of steps. For each image of such sequence, we apply a set of pretrained networks that predict the following: a) the location of the face (bounding box), using [38], b) an identity score for each image of the sequence that expresses the similarity between the original image (central image of the sequence) and each of the rest, using ArcFace [6], c) an age, race, and gender score using FairFace [16], d) a set of CelebA attributes classifiers (e.g., smile, wavy hair, etc.), and e) an estimation of the face pose (yaw, pitch, roll), using Hopenet [8]. In this way, for each warping we have a set of paths in the latent space and the corresponding paths in the attribute space.

In order to obtain a measure on how well the paths generated by a warping function are correlated with a certain attribute, we estimate the average Pearson's correlation between the index of the step along the path and the corresponding values in the attribute vector. By doing so, for each warping, we obtain a vector, which we normalize. This allows for sorting the discovered paths with respect to the correlation with each attribute and select the paths that give the maximum absolute correlation for each attribute.

The results are summarised in Table 2, where we report quantitative results for our method (Tab. 2a), in comparison to [34] (Tab. 2b) and [11] (Tab. 2c), in terms of $\mathcal{L}_1$-normalized correlation averaged across 100 latent codes. We note that our method achieves better correlations for the respective attributes, while at the same time the correlations with the rest of the attributes are lower than those achieved by [34, 11], as is evident by the lower values in the off-diagonal elements of the matrix. This shows in a quantitative manner, what was evident in a qualitatively manner in Fig. 8, that is, that the discovered paths in the latent space lead to more disentangled changes in the attribute space.

Finally, in Fig. 5 we show the results of generation across some non-linear interpretable paths obtained automatically by our method for StyleGAN2, for the following attributes: *age*, *race* (skin color), *gender* ("femaleness"), and *yaw* (rotation). In this figure, we report the paths with the highest correlation with the respective attribute.

5. Conclusion

In this paper, we presented our method for discovering non-linear interpretable paths in the latent space of pretrained GANs in an unsupervised and model-agnostic manner. We do so by modeling non-linear latent paths using the gradient of RBF-based warping functions, which we optimized in order to be distinguishable to each other. This leads to paths that correspond to interpretable generation where only a small number of generative factors are affected for each path. Finally, we proposed a quantitative evaluation protocol for the case of face-generating GANs, which can be used to automatically associate the discovered paths with interpretable attributes such as smiling and rotation.

Acknowledgments: This work was supported by the EU H2020 AI4Media No. 951911 project.

References

[1] D. Bau, J. Zhu, H. Strobelt, B. Zhou, J. B. Tenenbaum, W. T. Freeman, and A. Torralba. GAN dissection: Visualizing and understanding generative adversarial networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019.

[2] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[3] A. Brock, J. Donahue, and K. Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019.

[4] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2172–2180, 2016.

[5] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009.

[6] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[7] E. Denton, B. Hutchinson, M. Mitchell, and T. Gebru. Detecting bias with generative counterfactual face attribute augmentation. *arXiv preprint arXiv:1906.06439*, 2019.

[8] B. Doosti, S. Naha, M. Mirbagheri, and D. J. Crandall. Hopenet: A graph-based model for hand-object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6608–6617, 2020.

[9] L. Goetschalckx, A. Andonian, A. Oliva, and P. Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5744–5753, 2019.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[11] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris. GANSpace: Discovering interpretable GAN controls. *CoRR*, abs/2004.02546, 2020.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[13] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

[14] A. Jahanian, L. Chai, and P. Isola. On the "steerability" of generative adversarial networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[15] Y. Jin, J. Zhang, M. Li, Y. Tian, and H. Zhu. Towards the high-quality anime characters generation with generative adversarial networks. In *Proceedings of the Machine Learning for Creativity and Design Workshop at NeurIPS*, 2017.

[16] K. Kärkkäinen and J. Joo. Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913*, 2019.

[17] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

[18] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[19] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8107–8116. IEEE, 2020.

[20] Y. LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[21] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[22] W. Lee, D. Kim, S. Hong, and H. Lee. High-fidelity synthesis with disentangled representation. *CoRR*, abs/2001.04296, 2020.

[23] B. Liu, Y. Zhu, Z. Fu, G. de Melo, and A. Elgammal. OOGAN: disentangling GAN with one-hot sampling and orthogonal regularization. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 4836–4843, 2020.

[24] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 3730–3738. IEEE Computer Society, 2015.

[25] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

[26] A. Plumerault, H. L. Borgne, and C. Hudelot. Controlling generative models with continuous factors of variations. In *8th International Conference on Learning Representa-*

tions, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, 2020.

[27] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[28] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In Y. Bengio and Y. LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[29] A. Ramesh, Y. Choi, and Y. LeCun. A spectral regularizer for unsupervised disentanglement. *arXiv preprint arXiv:1812.01161*, 2018.

[30] Y. Shen, J. Gu, X. Tang, and B. Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9243–9252, 2020.

[31] Y. Shen, J. Gu, X. Tang, and B. Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9243–9252, 2020.

[32] N. Spingarn, R. Banner, and T. Michaeli. GAN ”steerability” without optimization. In *International Conference on Learning Representations*, 2021.

[33] A. Voynov and A. Babenko. RPGAN: gans interpretability via random routing. *CoRR*, abs/1912.10920, 2019.

[34] A. Voynov and A. Babenko. Unsupervised discovery of interpretable directions in the GAN latent space. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9786–9796. PMLR, 2020.

[35] T. Xiao, J. Hong, and J. Ma. Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 168–184, 2018.

[36] X. Xing, T. Han, R. Gao, S.-C. Zhu, and Y. N. Wu. Unsupervised disentangling of appearance and geometry by deformable generator network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10354–10363, 2019.

[37] C. Yang, Y. Shen, and B. Zhou. Semantic hierarchy emerges in deep generative representations for scene synthesis. *CoRR*, abs/1911.09267, 2019.

[38] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li. S3fd: Single shot scale-invariant face detector. In *Proceedings of the IEEE international conference on computer vision*, pages 192–201, 2017.