Your face can be changed in a variety of ways

Changyeon Yoon
Seoul National University
shinypond@snu.ac.kr

Yeonsu Lee
Seoul National University
dldustn990@snu.ac.kr

Jaemoo Choi
Seoul National University
toony42@snu.ac.kr

Abstract

Despite generating high-quality images, traditional Generative Adversarial Networks (GANs) have shown a lack of modifying already generated images. However, according to recent works, it is feasible to attain a set of interpretable directions for image manipulation in the latent space of the style-based GAN architecture (StyleGAN) under the supervised or unsupervised manner. In particular, some state-of-the-art models, such as GANSpace and StyleCLIP, manipulate a human facial image very naturally along the desired direction. Nevertheless, such approaches mainly assume a significant variation of a face, such as a change of gender or age, happens along only a one-dimensional axis as other typical face attributes. Furthermore, their proposed direction is global across the entire latent space. Unfortunately, such linear change of the latent vector yields escape from the manifold it belonged to and significantly degrades the quality of the generated image. To address these issues, we propose a method to output different manipulation results for a given semantic without escaping from the latent manifold. This approach allows us to obtain much more diverse and high-quality facial images compared to existing state-of-the-art models.

1. Introduction

Modern Generative Adversarial Networks (GANs) [8], like ProGAN [12], BigGAN [3], StyleGANs [14, 15, 13], have shown remarkable abilities to synthesize a variety of high-fidelity images. Unfortunately, despite the high quality of the output images, it is still challenging to modify a generated sample in the desired direction by adjusting its latent vector while preserving its quality. Several approaches [18, 9, 19, 16] interpret which latent direction represents a meaningful semantic part and thus make the user control for image manipulation possible. In particular, StyleCLIP [16], a state-of-the-art text-driven manipulation model, shows outstanding results in facial image manipulation and introduces many attractive directions to be tuned, such as gender, age, pose, hairstyle, and a specific person's style.

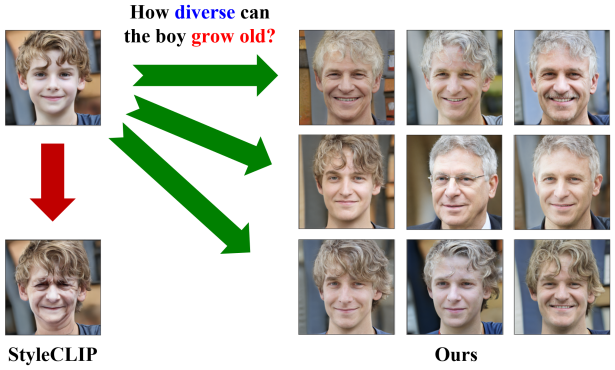

Figure 1. The main concept of our work. We propose a method to manipulate a given facial image while obtaining both diversity and high quality simultaneously, compared to StyleCLIP, a state-of-the-art text-driven image manipulation model.

However, there are two major problems with these image manipulation methods. One problem is that some semantic attributes, such as age and gender, have the potential to vary enormously, and existing technique often treat them as if they were one-dimensional. In general, any given facial image can grow old or be changed to another gender in different ways, not along a single axis, as shown in Figure 1. Another problem is that when modifying a particular image, previous methods just move the latent vector only linearly along the global direction [18, 9, 16]. The use of a global direction is convenient in that it reduces the inference time, but it degrades the quality of image manipulation at some local points since it is commonly applied to all points in the latent manifold. Figure 2 shows that a manipulation along the global direction eventually leads to extremely poor quality. Due to these two problems, previous approaches provide only little insight in exploring the high-dimensional image manifold derived from GANs.

In this paper, we address these issues by proposing a method to output different manipulation results for a given semantic without escaping from the latent manifold. Indeed, we apply stochastic iterative traversal with a global

Figure 2. High intensity of image manipulation along a global direction eventually leads to extremely poor quality. This phenomenon can be interpreted as the latent vector escaping the latent manifold $\mathcal{W}^+$. In this figure, we adopt a global direction corresponding to the text prompt 'old'. For the implementation details, see Section 5.1.

direction from StyleCLIP [16], based on the iterative curve-traversal introduced in [5]. The main concept is substituting the global direction to a local basis vector, which is the most similar to the global direction. Here, the local basis is obtained from the singular value decomposition (SVD) of the Jacobian matrix of the mapping network in StyleGAN [5]. This technique guarantees the latent vector not to escape from the latent manifold $\mathcal{W}^+$ so that the quality of generated image is always preserved during the traversal. Moreover, we further increase the diversity of generated images by introducing randomness of the step size for each iteration. Figure 3 and Figure 4 show relatively diverse image manipulation compared to StyleCLIP [16].

Our contributions are summarized as the followings:

- We propose a method to obtain both diversity and high quality of manipulated images for a coarse-level semantic attribute.
- We identify that image manipulation along the global direction makes the latent vector escape from the manifold and yields a poor-quality result.

2. Related Works

Style-based generators In recent years, GANs equipped with style-based generators [14, 15] have shown state-of-the-art performance in terms of high-fidelity image synthesis. The style-based generator consists of two parts: a mapping network, which encodes the initial latent code $z \in \mathcal{Z}$ to the style codes $w \in \mathcal{W}$, and a synthesis network, which takes the style codes w as the input and yields an image as the output. Specifically, StyleGAN [14] uses the style codes to control channel-wise means and variances through Adaptive Instance Normalization (AdaIN) [10], while StyleGAN2 [15] uses the style codes to control channel-wise variances by modulating the weights of each convolution layer. Despite the considerable improvement of image quality, however, there lacks enough understanding about the influence of a slight movement in the warped intermediate latent space $\mathcal{W}$ [14] on the image space. Thus, the need for research to analyze and control the latent space of the style-based generators emerges.

Latent Semantic Interpretation A large number of GAN models have a high potential for learning semantic factors from data. [4] add a regularization term to learn an interpretable factorized representation. [2] encode a variety of semantic factors in the intermediate feature space. Also, [11, 7, 23, 18] enable user control over the semantic attributes of the output under supervised learning of latent directions. Unlike these methods that work in the supervised manner, [21] propose an unsupervised optimization method to jointly learn a candidate matrix and a corresponding reconstructor, which identifies the semantic direction in the matrix. [9] find a global basis of $\mathcal{W}$ for latent space control using Principal Component Analysis (PCA), without any labels of output images. [19] propose a closed-form factorization of latent semantics without any sampling or additional training, while [9] require a large number of random sampling of latent vectors. [16] achieve state-of-the-art performance in text-driven image manipulation of StyleGAN imagery, dealing with three different points of view: latent optimization, latent mapper, and global directions. Meanwhile, most of the recent works focus on a controllable change along a fixed semantic direction. In other words, when a facial attribute that we want to control is determined, previous methods manipulate the corresponding semantic part as if it is one-dimensional. Even though [9] has a stochastic property due to Monte-Carlo sampling of latent vectors, the set of global directions obtained from the induced PCA basis is fixed during inference. In this aspect, our proposed methods aim to vary the results of facial image manipulation when the target attribute is one of the fundamental features of the human face, like gender, age, or emotions, primarily based on the architectures of StyleCLIP [16]. To the best of our knowledge, there has been no approach adopted this kind of viewpoint.

Latent Optimization One of the simple approaches to guide image manipulation is optimizing the latent $w \in \mathcal{W}^+$ with an adequate loss. To obtain a manipulated image containing the desired semantic, StyleCLIP [16] introduce two novel methods, the Latent Optimization and the Latent Mapper, while considering facial consistency as well as targeting features. Given a source latent code $w_s \in \mathcal{W}^+$ and a text prompt t, the Latent Optimization has an objective function as the following:

$$\arg\min_{w \in \mathcal{W}^+} L(G(w), t) + \lambda_w \|w - w_s\|^2 + \lambda_{\mathrm{ID}} L_{\mathrm{ID}}(w), \quad (1)$$

where L is a CLIP [17] loss and G is a pre-trained StyleGAN2 generator. Note that the consistency of generated image is forced by the L^2-regularization in the latent space $\mathcal{W}$ and the following identity loss:

$$L_{\mathrm{ID}}(w) = 1 - \langle R(G(w_s)), R(G(w)) \rangle, \quad (2)$$

where R is a pre-trained ArcFace [6] network for face recognition. Unfortunately, this method is highly sensitive to hyperparameters λ_w and λ_{ID}, indicating it is not practical. To address this issue, StyleCLIP also suggest the Latent Mapper, derived from the Latent Optimization, to further learn the change of semantics at each latent vector in $\mathcal{W}^+$. However, despite their remarkable performance, the resultant mapper networks are all distinct for each image and each text prompt. Thus, this method has a disadvantage that the training time is too long (about 10 - 12 hours per a single image and a single text, according to Table 1 in [16]). Our proposed method is similar to these optimization techniques in that it follows the desired direction in the latent space. Nevertheless, our work does not depend on a specific loss, such as (1), which can limit the diversity of results. Instead, we maintain the image quality by trying to keep the latent vector from leaving the manifold, and significantly improve the execution time by exploiting only closed-form computations such as SVD.

3. Background

Our work is primarily based on StyleCLIP [16], which show a state-of-the-art performance in text-driven image manipulation. In StyleCLIP, there are two representative manipulation methodologies; one to solve the optimization problem for each image, and the other to find a unique global direction for each desired property. These methods are complementary to each other and have some drawbacks. Our approach mainly improves these drawbacks while generating various images on one feature.

3.1. Global Direction

One method proposed in [16] aims to find a global direction Δs which indicates the change of a specific trait given by a text prompt. Under the assumption of the existence and the uniqueness of the global direction Δs in the style space $\mathcal{S}$ [22], the objective is to find Δs that $G(s + \alpha\Delta s)$ represents the semantic attribute stronger than $G(s)$, where $\alpha > 0$ is a hyperparameter.

Explicitly, we start from a text embedding. Let Δt be the difference between a neutral text embedding and a target text embedding. For instance, when manipulating a person's face older, a target text can be written as 'an elderly person' or 'an old person' or 'an old man' in each case. Here, the neutral class for each text prompt can be defined by 'a person' or 'a man.' Then we obtain Δt by subtracting the embedded vector of the neural text from the embedded vector of the target text. Additionally, prompt engineering is also required, as shown in [16]. Similarly, we define $i + \Delta i$ and i by the embedded vectors of $G(s + \Delta s)$ and $G(s)$, respectively. Finally, the goal is to find a vector Δs such that the induced Δi highly correlates to Δt already obtained.

Now, let Δs_c be zero except the c-th component, which is set to the standard deviation of the channel from 100 image samples. By defining Δi_c as the difference between the embedded vectors of $G(s + \alpha\Delta s_c)$ and $G(s - \alpha\Delta s_c)$, we measure

$$R_c(\Delta i) = \mathbb{E}_{s\in\mathcal{S}}[\Delta i_c \cdot \Delta t], \tag{3}$$

where R_c stands for the relevance of the channel c with the attribute and α is set to 5, which is the magnitude of the perturbation. Next, we apply a threshold β. If $|R_c|$ does not exceed β, then we regard the channel c as a negligible part, i.e, having no relationship with the attribute Δt. Finally, we define Δs by:

$$[\Delta s]_c = \begin{cases} R_c & \text{if } |R_c| \geq \beta, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

[16] empirically show that the global direction shows better performance than the optimization method. Nonetheless, the global direction always pushes all latent vectors along only one direction, contrary to the optimization method which considers the local properties of the latent space. This causes two major problems: in terms of diversity and quality. First, the global direction severely deteriorates the diversity of manipulated samples. For instance, when manipulating a person older, it eventually makes the similar old person even if the starting images are different, since 'the old man' corresponds to a unique global direction. Furthermore, as shown in Figure 2, high intensity of manipulation along a global direction yields poor-quality images since the global direction hardly considers the local property. On the other hand, our proposed method acquires both diversity and quality of manipulated images; because it embraces the global direction when finding the most similar vector while exploiting the advantages of the optimization method.

3.2. Subspace Traversal

One of the main concepts of StyleGAN is the following mapping network:

$$f : \mathcal{Z} \to \mathcal{W}, \tag{5}$$

where $\mathcal{Z}$ is the latent space with a Gaussian distribution, $\mathcal{W}$ is the intermediate latent space, and f is an eight-layered-MLP. This mapping network enables to feed of disentangled and refined vectors as a new input to the StyleGAN generator. However, not all elements of $\mathcal{W}$ generate what we intended, and several images generated from arbitrary latent vectors in $\mathcal{W}$ are even unrecognizable. Hence, let $\mathcal{I}$ be a subspace of $\mathcal{W}$ that yields photo-realistic images. [5] observe that the necessary condition for a latent vector in $\mathcal{W}$ to belong to $\mathcal{I}$ is that the vector should be in the range of $f : \mathcal{Z} \to \mathcal{W}$. In other words, the following identity holds:

$$\mathcal{I} \subseteq f(\mathcal{Z}) \subsetneq \mathcal{W}. \tag{6}$$

(a) (b)

(c) (d)

Figure 3. Comparison of StyleCLIP and our proposed method on a change of age: (a) linear traversal along the global direction 'old' from StyleCLIP, (b) iterative traversal along our local basis vector, which is the most similar to the global direction 'old' from StyleCLIP, (c) linear traversal along the opposite global direction 'old' from StyleCLIP, (d) iterative traversal along our local basis vector, which is the most similar to the opposite global direction 'old' from StyleCLIP. Note that the old man in the first row of (c) and (d) is from the last row of the seventh column in (b).

For further analysis, assume that $\mathcal{I}$ is a k-submanifold embedded in $\mathcal{W}$. Then our objective is to force the latent vector to traverse in manifold $\mathcal{I}$. A simple but efficient method is to restrict $\mathcal{W}$ into k-dimensional subspace locally. First, let

$$J(z) = \frac{\partial f}{\partial z}(z), \tag{7}$$

and denote the tangent space of $\mathcal{I}$ by

$$T_w\mathcal{W} = \{J(z)x : x \in \mathbb{R}^m\}, \tag{8}$$

where $w = f(z)$, and m is the dimension of $\mathcal{Z}$. Note that the most actively changing k-subspace is the (locally) best approximation of $\mathcal{I}$.

Finally, to obtain a proper basis for k-subspace $\mathcal{I}$ in $\mathcal{W}$, the singular value decomposition (SVD) of (7) can be exploited as the following:

$$J(z) = U\Sigma V^T, \tag{9}$$

where U and V are orthogonal matrices, and Σ is a diagonal matrix consisting of singular values. Here, the columns of U form a basis of $\mathcal{W}$, and the columns of V form a basis of $\mathcal{Z}$. In particular, the columns of U which correspond to the top k singular values are exactly the desired local coordinate of $\mathcal{W}$ at $w \in \mathcal{I}$.

4. Methods

In this section, we develop Section 3 to obtain photo-realistic manipulated images. First, we naturally expand the discussion in Section 3.2 to $\mathcal{W}^+$ space [14, 1]. Next, we combine this with the global direction introduced in Section 3.1 to obtain diverse images.

4.1. Subspace Traversal in $\mathcal{W}^+$

First of all, note that we generally invert an image to a vector $\tilde{w}$ in $\mathcal{W}^+$ space [20], which has 18 style vectors of dimension m. Unfortunately, (7) cannot be computed in this case since $z \in \mathcal{Z}$ corresponding to $\tilde{w}$ is unknown. To address this issue, let $w_1, w_2, \cdots, w_{18}$ be the style vectors, each fed into different stages of the generator. Then $\tilde{w}$ is written as

$$\tilde{w} = (w_1, w_2, \ldots, w_{18}). \tag{10}$$

For each vector w_i, the corresponding $z_i = g(w_i)$ is obtained by optimizing a simple inversion MLP g with a loss function given by

$$L_{\text{inv}}(w_i) = \|f(g(w_i)) - w_i\|_2^2. \tag{11}$$

Therefore, by letting $\tilde{z} = (z_1, z_2, \ldots, z_2)$ and $\tilde{f}(\tilde{z}) = (f(z_1), f(z_2), \ldots, f(z_{18}))$, we gain

$$\tilde{J} = \partial\tilde{f}/\partial\tilde{z}. \tag{12}$$

Regarding $\tilde{J}$ as the counterpart of J in (7), the remaining part of the discussion is exactly the same as Section 3.2.

Now, let v be a global direction obtained by the method introduced in Section 3.1. Also, let P be a k-dimensional tangent space induced from $\tilde{J}$. Then we adopt the most similar orientation to the global direction among the basis vectors, obtained through SVD. The chosen orientation can be interpreted as the first principal component in the Principal Component Analysis (PCA) at the local point. We empirically identify that this technique can manipulate a given image, while preserving the image quality. (see Figure 3 and Figure 4).

4.2. Diverse Manipulation

Here, we demonstrate how to achieve the high diversity of results. Note that such discussion is essential since a single attribute does not imply that it has only one direction. Indeed, one can get older, or be angrier, or be cuter in various ways. Therefore, with the subspace traversal suggested in Section 4.1, we additionally suggest some techniques to increase the diversity of manipulated images.

Random Step Size The first method is adjusting the step size randomly for each iteration. As proposed in Section 4.1, the traversal path should be in the local k-subspace, which is spanned by the selected basis. Thus, as the step size varies randomly, the local basis induced by the latent vector slightly changes, and this phenomenon can produce a variety of progressive paths. For the implementation details of randomness in the step size, see Section 5.1.

Overlapping each layer As mentioned in Section 4.1, after inverting 18 blocks of each $\mathcal{W}^+$ into $\mathcal{Z}$, the Jacobian $J(z)$ of the mapper network f for each block is computed and its singular value decomposition yields a local k-dimensional basis. Then, one element of basis with the largest inner product value for the given global direction is selected as the next direction of progress. However, the chosen vectors are belonged to $\mathcal{W}$ so that the actual step size interpreted in $\mathcal{Z}$ highly depends on the corresponding singular values. To address this issue, one of the simplest methods is moving the latent vector by the same unit distance in $\mathcal{Z}$, not $\mathcal{W}$ or $\mathcal{W}^+$. Explicitly, we pull-back each vector to $T_z\mathcal{Z}$ and simply accumulate these vectors as the to restrict into one $T_z\mathcal{Z}$ space. Note that we add all these vectors rather than concatenation. By such addition, the update direction of each block in $\mathcal{W}^+$ overlaps each other, and this method further stimulates to increase the variety of our image manipulation results. Finally, we push forward the tangent vector from $T_z\mathcal{Z}$ to $T_w\mathcal{W}^+$ with the randomized step size, to generate a manipulated image.

(a) (b)

(c) (d)

Figure 4. Comparison of StyleCLIP and our proposed method on a change of gender: (a) linear traversal along the opposite global direction ‘man’ from StyleCLIP, (b) iterative traversal along our local basis vector, which is the most similar to the opposite global direction ‘man’ from StyleCLIP, (c) linear traversal along the global direction ‘man’ from StyleCLIP, (d) iterative traversal along our local basis vector, which is the most similar to the global direction ‘man’ from StyleCLIP. Note that the woman in the first row of (c) and (d) is from the last row of the nineth column in (b).

Algorithm 1 Stochastic Iterative Traversal

Input: A latent vector $w \in \mathcal{W}^+ = \mathbb{R}^{m \times 18}$, a global direction $v \in \mathcal{W}^+$ corresponding to a text t, the subspace dimension $k \in [1, m]$, and the step size hyperparameters a, b.

for iteration **do**
  Find z corresponding to w
  $J(z) \leftarrow \left(\partial f / \partial z\right)(z)$
  Obtain U, V from SVD of $J(z) = U\Sigma V^T$
  $i \leftarrow \arg\max\{|U_1^T v|, |U_2^T v|, \ldots, |U_k^T v|\}$
  $v' \leftarrow \text{sign}\left(U_i^T v\right)\ V_i^T v$
  $w \leftarrow w + \lambda v'$ where $\lambda \sim U_{[a,b]}$
end for
return $G(w)$

5. Experiments

5.1. Implementation Details

Finding a global direction We follow the implementation of [16], which is also explained in Section 3.1. The only difference with [16] is that we experiment on $\mathcal{W}^+$, the latent of StyleGAN, rather than StyleSpace $\mathcal{S}$ [22]. We used "old person" for the target attribute and "person" for the neutral class in manipulating age. Similarly, we apply "man" for the target attribute and "a person" for the neutral class in manipulating gender. Next, after the aggregation of 2824 latent codes $\tilde{w} \in \mathcal{W}^+$, we calculated standard deviations of every channels for perturbation. As $R_c(\Delta i)$ is the expectation of dot product of Δi_c and Δi (3), we can use the same Δi_c for different Δi. If we computed Δi_c and save them first, then we can reuse them for every time we calculate $R_c(\Delta i)$ for different Δi which is the same as Δt. So we used 42 image pairs to compute 42 Δi_c and the mean of them. The images are generated images from latent vectors $\tilde{w} \pm \alpha \Delta \tilde{w}_c$, where $\Delta \tilde{w}_s$ is a zero vector, except its c coordinate set to the standard deviation of that channel we computed above. Initially, we tried to use 100 image pairs, but it seemed that it would take too long to have results. Therefore we used the largest batch size available to us, which was 14, and iterated 3 times to get the mean values. These 42 latent codes were obtained from the collection, which has inverted latent codes of CelebA-HQ. It took about 9 hours to get 42 Δi_c for all 9216 channels. Once we have the mean of Δi_c, we can simply get the relevance of channel c to the target manipulation $R_c(\Delta i)$ by dot product of the mean of Δi_c and Δi. We set $\alpha = 5$ and $\beta = 0.14$ initially, but if the obtained global direction was too sparse, we lowered β so that global direction had between 50 and 150 non-zero channels. Specifically, we used $\beta = 0.14$ in manipulating age, $\beta = 0.09$ in manipulating gender.

Stochastic Traversal Figure 3 (a) and (c) are obtained by increasing the step size by 2. Figure 3 (b) changes the step size uniformly at random from 0.05 to 0.20 with iteration 30. Figure 3 (d) changes the step size uniformly at random from 0.10 to 0.25 with iteration 30. All images in Figure 3 (b) and (d) are drawn for every six iterations. Figure 4 (a) and (c) are obtained by increasing the step size by 2. Figure 4 (b) changes the step size uniformly at random from 0.05 to 0.12 with iteration 30. Figure 4 (d) changes the step size uniformly at random from 0.10 to 0.18 with iteration 30. All images in Figure 4 (b) and (d) are drawn for every six iterations.

5.2. Results

The results of the global direction and our method are demonstrated in Figure 3 and Figure 4. Images shown in Figure 3 (a) present manipulated image by the global direction of attribute of 'aging'. (c) shows the result by traversing in the opposite direction. The top image is the original image, and the others are manipulated images with different intensity (α). Similarly, Figure 4 (a) and (c) show the manipulated results by the global directions, which each direction stands attribute of 'man' and 'woman', respectively. Moreover, (b) and (d) of these two figures demonstrate our diverse outcome where (b) and (d) are counterpart of (a) and (c) of each figures, respectively.

Quality As shown in Figure 2, if the step size α of the global direction is large, the image collapses to unknown shape. Furthermore, in Figure 3 (a) and (c), the modified version of the attribute 'elderly' and 'youthful' look malformed. Even Figure 4 (a) does not change the gender, indicating another failure of the global direction. In contrast, we observe that the image does not collapse even after many iterations when using our proposed method. As shown in Figure 3 and Figure 4, the stochastic iterative traversal has the potential to keep the latent vector from leaving the image manifold. In conclusion, the image manipulation task was successful in terms of the image quality.

Diversity We show that with Figure 3 and Figure 4, various images can be retrieved in one attribute. Although the global direction is the same, the manipulation results are varied by restricting the direction to local image manifold and randomizing step size. Obviously, tremendous changes are demanded to manipulate age and gender, which forces to go through a long journey. Shown in Figure 3 and Figure 4, while global direction method experience despair throughout the journey, every pathways successfully concludes their expedition in variety ways within our method.

6. Conclusion

Many recent studies in facial image manipulation have treated some coarse-level attributes, such as age and gender, as if they were one-dimensional. Furthermore, their proposed global directions often degrade the image quality, as the intensity of manipulation increases. On the contrary, our method enables a long, stable traversal by restricting the update direction into a submanifold of $\mathcal{W}^+$ induced from the mapping network of StyleGAN. By comprehensive experiments, we demonstrates our traversal does not escape from the latent manifold and preserves the quality of manipulated images. Also, we show that even the same attribute can yield different results by adding randomness of the step size for each iteration. Hence, our proposed methods allow us to obtain much more diverse and high-quality facial images compared to existing state-of-the-art manipulation models, such as StyleCLIP.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019. 5

[2] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B Tenenbaum, William T Freeman, and Antonio Torralba. Gan dissection: Visualizing and understanding generative adversarial networks. *arXiv preprint arXiv:1811.10597*, 2018. 2

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 1

[4] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *arXiv preprint arXiv:1606.03657*, 2016. 2

[5] Jaewoong Choi, Changyeon Yoon, Junho Lee, Jung Ho Park, Geonho Hwang, and Myungjoo Kang. Do not escape from the manifold: Discovering the local coordinates on the latent space of gans. *arXiv*, 2021. 2, 3

[6] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos 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. 3

[7] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5744–5753, 2019. 2

[8] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014. 1

[9] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *arXiv preprint arXiv:2004.02546*, 2020. 1, 2

[10] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 2

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

[12] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 1

[13] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *arXiv preprint arXiv:2006.06676*, 2020. 1

[14] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 1, 2, 5

[15] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 1, 2

[16] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. *arXiv preprint arXiv:2103.17249*, 2021. 1, 2, 3, 7

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 2

[18] 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*, pages 9243–9252, 2020. 1, 2

[19] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. *arXiv preprint arXiv:2007.06600*, 2020. 1, 2

[20] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *arXiv preprint arXiv:2102.02766*, 2021. 5

[21] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In *International Conference on Machine Learning*, pages 9786–9796. PMLR, 2020. 2

[22] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. *arXiv preprint arXiv:2011.12799*, 2020. 3, 7

[23] Ceyuan Yang, Yujun Shen, and Bolei Zhou. Semantic hierarchy emerges in deep generative representations for scene synthesis. *International Journal of Computer Vision*, pages 1–16, 2021. 2