Interpreting the Latent Space of GANs for Semantic Face Editing

Yujun Shen[1], Jinjin Gu[2], Xiaoou Tang[1], Bolei Zhou[1]
[1]The Chinese University of Hong Kong [2]The Chinese University of Hong Kong, Shenzhen
{sy116, xtang, bzhou}@ie.cuhk.edu.hk, jinjingu@link.cuhk.edu.cn

Figure 1: Manipulating various facial attributes through varying the latent codes of a well-trained GAN model. The first column shows the original synthesis from PGGAN [17], while each of the other columns shows the results of manipulating a specific attribute.

Abstract

Despite the recent advance of Generative Adversarial Networks (GANs) in high-fidelity image synthesis, there lacks enough understanding of how GANs are able to map a randomly sampled latent code to a photo-realistic image. In this work, we propose InterFaceGAN *for semantic face editing by interpreting the latent semantics learned by GANs. In this framework, we conduct a detailed study on how different semantics are encoded in the latent space of GANs for face synthesis. We find that the latent code of well-trained generative models actually learns a disentangled representation after linear transformations. We explore the disentanglement between various semantics and manage to decouple some entangled semantics with subspace projection, leading to more precise control of facial attributes. Besides manipulating gender, age, expression, and the presence of eyeglasses, we can even vary the face pose as well as fix the artifacts accidentally generated by GAN models. The proposed method is further applied to achieve real image manipulation when combined with GAN inversion methods or some encoder-involved models.*[1]

[1]Codes are available at https://genforce.github.io/interfacegan/.

1. Introduction

Generative Adversarial Networks (GANs) [13] have significantly advanced image synthesis in recent years. However, few efforts have been made on studying what a GAN actually learns with respect to the latent space. In this paper, we propose a framework *InterFaceGAN*, short for *Inter*preting *Face GAN*s, to identify the semantics encoded in the latent space of well-trained face synthesis models and then utilize them for semantic face editing. This framework provides both theoretical analysis and experimental results to verify that *linear* subspaces align with different *true-or-false* semantics emerging in the latent space. We further study the disentanglement between different semantics and show that we can decouple some entangled attributes (*e.g.*, old people are more likely to wear eyeglasses then young people) through the linear subspace projection. These disentangled semantics enable precise control of facial attributes with any given GAN model *without retraining*. Besides gender, age, expression, and the presence of eyeglasses, we can noticeably also vary the face pose or correct some artifacts produced by GANs. Some results are shown in Fig.1. Finally, we extend InterFaceGAN to real image manipulation through GAN inversion approaches.

Figure 2: Pose manipulation results by InterFaceGAN. Centered image is the original synthesis from PGGAN [17] model, while other images show the gradually changed poses.

1.1. Related Work

Generative Adversarial Networks. GAN [13] has brought wide attention in recent years due to its great potential in producing photo-realistic images [1, 15, 5, 34, 23, 17, 6, 18]. To make GANs applicable for real image processing, existing methods proposed to invert the generation process [25, 37, 22, 4, 14] or learn an additional encoder associated with the GAN training [11, 10, 36].

Study on Latent Space of GANs. Latent space of GANs is generally treated as Riemannian manifold [7, 2, 19, 20, 27]. Some work has observed the vector arithmetic property [26], but the study on how a well-trained GAN is able to encode different semantics inside the latent space is still missing. Some concurrent work also explores the latent semantics learned by GANs. Jahanian *et al.* [16] studies the steerability of GANs concerning camera motion and image color tone. Goetschalckx *et al.* [12] improves the memorability of the output image. Yang *et al.* [32] explores the hierarchical semantics in the deep generative representations for scene synthesis. Unlike them, we focus on facial attributes emerging in GANs for face synthesis and extend our method to real image manipulation.

Semantic Face Editing with GANs. Semantic face editing aims at manipulating facial attributes of a given image. To achieve this goal, existing approaches typically learned new models [24, 8, 30, 21, 33, 3, 31, 29, 9, 28]. Unlike these learning-based methods, this work explores the interpretable semantics inside the latent space of *fixed* GANs, and *turns unconstrained GANs to controllable GANs* by varying the latent code.

2. Framework of InterFaceGAN

In this section, we introduce the framework of InterFaceGAN, which provides a rigorous analysis of the semantics emerging in the latent space of GANs, and further leverages these semantics for facial editing.

Problem Statement. Given a well-trained GAN model, the generator can be formulated as a deterministic function $g : \mathcal{Z} \to \mathcal{X}$. Here, $\mathcal{Z} \subseteq \mathbb{R}^d$ denotes the d-dimensional latent space, for which Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ is commonly used [23, 17, 6, 18]. $\mathcal{X}$ stands for the image space, where each sample $\mathbf{x}$ possesses certain semantic information, like gender and age for face model. Suppose we have a semantic scoring function $f_S : \mathcal{X} \to \mathcal{S}$, where $\mathcal{S} \subseteq \mathbb{R}^m$ represents the semantic space with m semantics.

Figure 3: Examples on fixing the artifacts that PGGAN [17] has generated. First row shows some bad generation results, while the following two rows present the gradually corrected synthesis by moving the latent codes along the positive "quality" direction.

We can bridge the latent space $\mathcal{Z}$ and the semantic space $\mathcal{S}$ with $\mathbf{s} = f_S(g(\mathbf{z}))$, where $\mathbf{s}$ and $\mathbf{z}$ denote the semantic scores and the sampled latent code respectively.

Semantic Boundary in the Latent Space. Any normal vector $\mathbf{n} \in \mathbb{R}^d$ in the latent space defines a hyperplane, which separates the entire space into two parts. This is very similar to the separation of binary facial attributes, such as male *v.s.* female. In particular, $\mathbf{n}$ defines a "distance" from a latent code $\mathbf{z}$ to this hyperplane as $\mathrm{d}(\mathbf{n}, \mathbf{z}) = \mathbf{n}^T\mathbf{z}$. Note that $\mathrm{d}(\cdot, \cdot)$ is not a rigorous distance since it can be negative. We expect this "distance" to be linearly dependent with the score of a particular semantic as $f(g(\mathbf{z})) = \lambda \mathrm{d}(\mathbf{n}, \mathbf{z})$. Here, $\lambda > 0$ is a scalar to measure how fast the semantic varies along this direction. Hence, the hyperplane defined by $\mathbf{n}$ serves as a semantic boundary in the latent space.

Semantic Manipulation in the Latent Space. With the normal vector $\mathbf{n}$ of the semantic boundary, we can edit the output image by varying the latent code $\mathbf{z}$ with $\mathbf{z}_{edit} = \mathbf{z} + \alpha\mathbf{n}$. It will make the synthesis look more positive on such semantic with $\alpha > 0$, since the score becomes $f(g(\mathbf{z}_{edit})) = f(g(\mathbf{z})) + \lambda\alpha$ after editing. Similarly, $\alpha < 0$ will make the synthesis look more negative.

Conditional Manipulation. When there is more than one attribute, editing one may affect another since some semantics can be coupled with each other. To achieve more precise control, we propose *conditional manipulation* by performing subspace projection. More concretely, given

Figure 4: Examples for conditional manipulation. The first two rows show the manipulation results along with the original directions learned by SVMs for two attributes independently. The last row edits the faces by varying one attribute with the other one unchanged.

two hyperplanes with normal vectors $\mathbf{n}_1$ and $\mathbf{n}_2$, we find a projected direction $\mathbf{n}_1 - (\mathbf{n}_1^T\mathbf{n}_2)\mathbf{n}_2$ such that moving samples along this new direction can change "attribute 1" without affecting "attribute 2". If there is more than one attribute to be conditioned on, we just subtract the projection from the primal direction onto the plane that is constructed by all conditioned directions.

Real Image Manipulation. To enable real image editing, we need to map a real image to a latent code. For this purpose, existing methods have proposed to directly optimize the latent code [22, 35] or to learn an extra encoder [37, 4]. There are also some models that have already involved an encoder along with the training process of GANs [11, 10, 36]. After getting the latent code, we can manipulate the target image as discussed above.

3. Experiments

In this section, we evaluate InterFaceGAN with PGGAN [17] and StyleGAN [18] on both synthesized and real faces.

3.1. Semantic Manipulation

Besides manipulating gender, age, expression, and the presence of eyeglasses, which is shown in Fig.1, we can produce gradually changed face pose (Fig.2) and even fix some artifacts accidentally generated by GANs (Fig.3). From Fig.2, we can tell that even there lacks enough data with extreme poses in the training data, *i.e.*, CelebA-HQ [17], GAN is capable of imagining how profile faces should look like. We can also conclude from Fig.3 that "quality" is also encoded in the latent space as a specific semantic.

3.2. Conditional Manipulation

To decorrelate different semantics for independent facial attribute editing, we propose conditional manipulation in Sec.2. Fig.4 shows some results by manipulating one attribute with another one as a condition. Taking the left sample in Fig.4 as an example, the results tend to become male when being edited to get old (top row). We fix this problem by subtracting its projection onto the gender direction from age direction, resulting in a new direction. In this way, we can make sure the gender component is barely affected when the sample is moved along the projected direction (bottom row). Fig.5 shows conditional manipulation with more than one constraint, where we add glasses by conditionally preserving age and gender. In the beginning, adding eyeglasses is entangled with changing both age and gender. But we manage to add glasses without affecting age and gender.

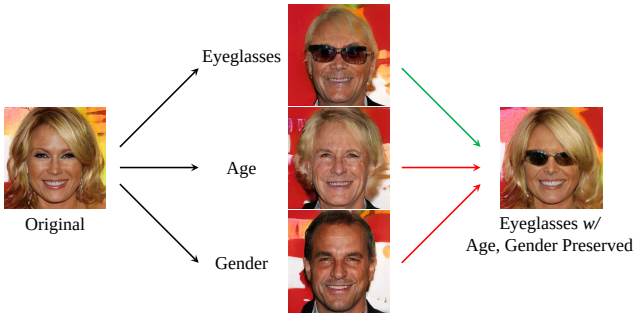

Figure 5: Examples for conditional manipulation with more than one conditions. Left: Original synthesis. Middle: Manipulations along single boundary. Right: Conditional manipulation. **Green** arrow: Primal direction. **Red** arrows: Projection subtraction.

3.3. Real Image Manipulation

In this part, we evaluate InterFaceGAN on real faces. Recall that InterFaceGAN achieves semantic face editing by moving the latent code along a certain direction. Hence, we need to first invert the given real image back to the latent code. To invert a pre-trained GAN model, there are two typical approaches. One is the optimization-based approach, which directly optimizes the latent code with the fixed generator to minimize the pixel-wise reconstruction error [22]. The other is the encoder-based, where an extra encoder network is trained to learn the inverse mapping [37]. We tested both of these two baseline approaches.

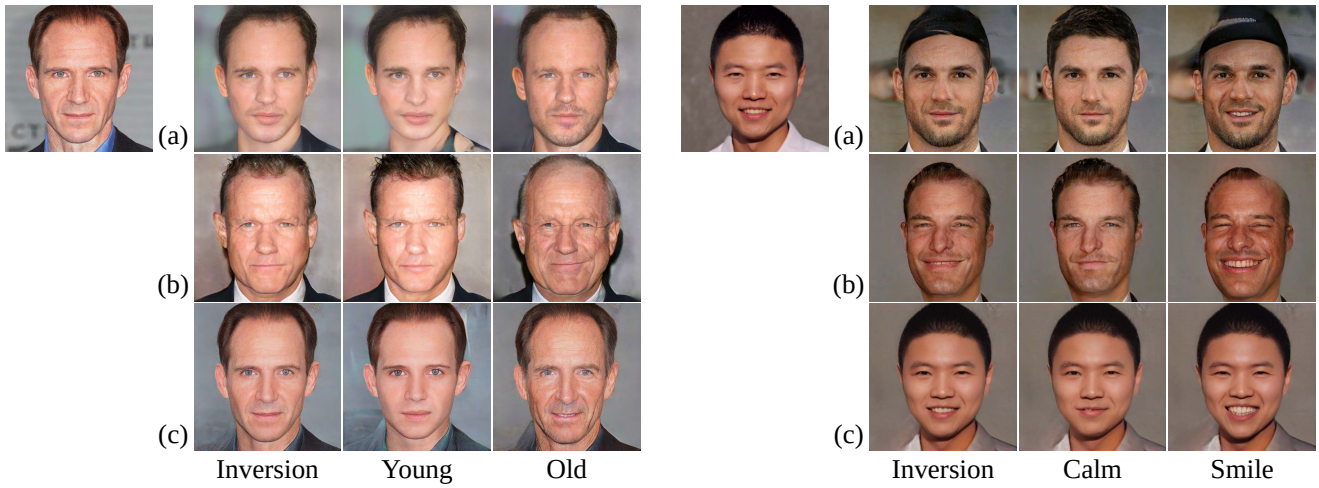

Figure 6: Manipulating real faces with respect to the attributes age and gender, using the pre-trained PGGAN [17] and StyleGAN [18]. Given an image to edit, we first invert it back to the latent code and then manipulate the latent code with InterFaceGAN. On the top left corner is the input real face. From top to bottom: (a) PGGAN with optimization-based inversion method, (b) PGGAN with encoder-based inversion method, (c) StyleGAN with optimization-based inversion method.

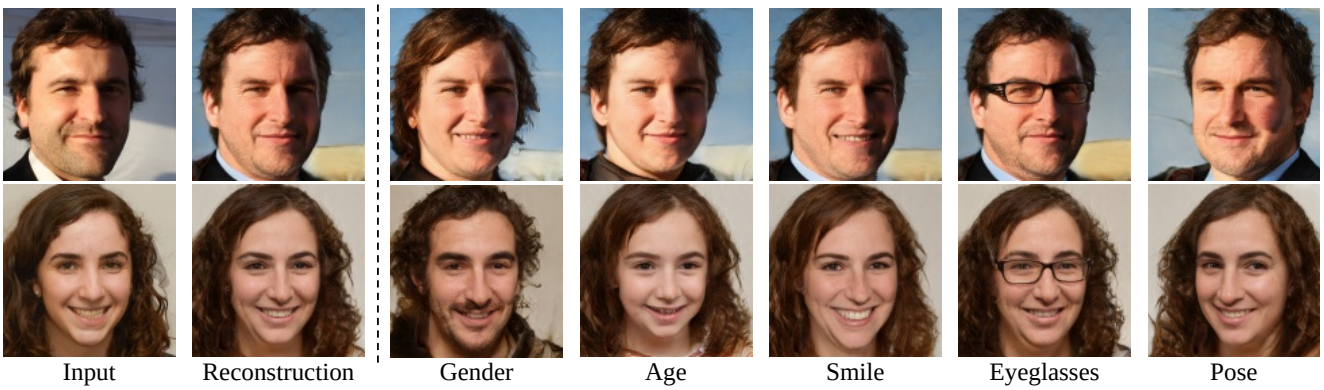

Figure 7: Manipulating real faces with LIA [36], which is a encoder-decoder generative model for high-resolution face synthesis.

Results are shown in Fig.6. We can tell that both optimization-based (first row) and encoder-based (second row) methods show poor performance when inverting PGGAN. This can be imputed to the strong discrepancy between training and testing data distributions. For example, the model tends to generate Western people even the input is an Easterner (see the right example in Fig.6). Even unlike the inputs, however, the inverted images can still be semantically edited with InterFaceGAN. Compared to PGGAN, the results on StyleGAN (third row) are much better. Here, we treat the layer-wise styles (*i.e.*, $\mathbf{w}$ for all layers) as the optimization target. When editing an instance, we push all style codes towards the same direction. As shown in Fig.6, we successfully change the attributes of real face images *without* retraining StyleGAN but leveraging the interpreted semantics from latent space.

We also test InterFaceGAN on encoder-decoder generative models, which train an encoder together with the generator and discriminator. After the model converges, the encoder can be directly used for inference to map a given image to latent space. We apply our method to interpret the latent space of the recent encoder-decoder model LIA [36]. The manipulation result is shown in Fig.7 where we successfully edit the input faces with various attributes, like age and face pose. It suggests that the latent code in the encoder-decoder based generative models also supports semantic manipulation. In addition, compared to Fig.6 (b) where the encoder is separately learned after the GAN model is well-prepared, the encoder trained together with the generator gives better reconstruction as well as manipulation results.

4. Conclusion

We propose InterFaceGAN to interpret the semantics encoded in the latent space of GANs. By leveraging the interpreted semantics as well as the proposed conditional manipulation technique, we are able to precisely control the facial attributes with any fixed GAN model, even turning unconditional GANs to controllable GANs.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017. 2

[2] Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent space oddity: on the curvature of deep generative models. In *ICLR*, 2018. 2

[3] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Towards open-set identity preserving face synthesis. In *CVPR*, 2018. 2

[4] David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. Seeing what a gan cannot generate. In *ICCV*, 2019. 2, 3

[5] David Berthelot, Thomas Schumm, and Luke Metz. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017. 2

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2019. 2

[7] Nutan Chen, Alexej Klushyn, Richard Kurle, Xueyan Jiang, Justin Bayer, and Patrick van der Smagt. Metrics for deep generative models. In *AISTAT*, 2018. 2

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

[9] Chris Donahue, Akshay Balsubramani, Julian McAuley, and Zachary C. Lipton. Semantically decomposing the latent spaces of generative adversarial networks. In *ICLR*, 2018. 2

[10] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. In *ICLR*, 2017. 2, 3

[11] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. In *ICLR*, 2017. 2, 3

[12] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *ICCV*, 2019. 2

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 1, 2

[14] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. In *CVPR*, 2020. 2

[15] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *NeurIPS*, 2017. 2

[16] Ali Jahanian, Lucy Chai, and Phillip Isola. On the "steerability" of generative adversarial networks. In *ICLR*, 2020. 2

[17] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018. 1, 2, 3, 4

[18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2, 3, 4

[19] Line Kuhnel, Tom Fletcher, Sarang Joshi, and Stefan Sommer. Latent space non-linear statistics. *arXiv preprint arXiv:1805.07632*, 2018. 2

[20] Samuli Laine. Feature-based metrics for exploring the latent space of generative models. In *ICLR Workshop*, 2018. 2

[21] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. Fader networks: Manipulating images by sliding attributes. In *NeurIPS*, 2017. 2

[22] Fangchang Ma, Ulas Ayaz, and Sertac Karaman. Invertibility of convolutional generative networks from partial measurements. In *NeurIPS*, 2018. 2, 3

[23] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *ICLR*, 2018. 2

[24] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *ICML*, 2017. 2

[25] Guim Perarnau, Joost Van De Weijer, Bogdan Raducanu, and Jose M Álvarez. Invertible conditional gans for image editing. In *NeurIPS Workshop*, 2016. 2

[26] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016. 2

[27] Hang Shao, Abhishek Kumar, and P Thomas Fletcher. The riemannian geometry of deep generative models. In *CVPR Workshop*, 2018. 2

[28] Yujun Shen, Ping Luo, Junjie Yan, Xiaogang Wang, and Xiaoou Tang. Faceid-gan: Learning a symmetry three-player gan for identity-preserving face synthesis. In *CVPR*, 2018. 2

[29] Yujun Shen, Bolei Zhou, Ping Luo, and Xiaoou Tang. Facefeat-gan: a two-stage approach for identity-preserving face synthesis. *arXiv preprint arXiv:1812.01288*, 2018. 2

[30] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, 2017. 2

[31] Taihong Xiao, Jiapeng Hong, and Jinwen Ma. Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. In *ECCV*, 2018. 2

[32] Ceyuan Yang, Yujun Shen, and Bolei Zhou. Semantic hierarchy emerges in deep generative representations for scene synthesis. *arXiv preprint arXiv:1911.09267*, 2019. 2

[33] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Towards large-pose face frontalization in the wild. In *ICCV*, 2017. 2

[34] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *ICML*, 2019. 2

[35] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. *arXiv preprint arXiv:2004.00049*, 2020. 3

[36] Jiapeng Zhu, Deli Zhao, and Bo Zhang. Lia: Latently invertible autoencoder with adversarial learning. *arXiv preprint arXiv:1906.08090*, 2019. 2, 3, 4

[37] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *ECCV*, 2016. 2, 3