Interpreting the Latent Space of GANs for Semantic Face Editing

Yujun Shen *et al.*
sy116@ie.cuhk.edu.hk
The Chinese University of Hong Kong

Along with extensive study on improving the synthesis quality of Generative Adversarial Networks (GANs), some notable study has also been done towards achieving controllable generation process by interpreting how different semantics are organized in the latent space of a well-trained GAN. Specifically, this paper focuses on revealing and manipulating various semantics encoded in the latent space of a GAN trained on facial dataset, in order to apply facial attribute editing. In contrast to previous work on semantic face editing, that mainly suggests special loss functions (e.g. [1]) and architectures (e.g. [4]), to get a disentangled representation of facial semantics and the identity information, the introduced framework, called InterFaceGAN, employs the idea of searching for navigation directions in the latent space to manipulate the generation process of a fixed pre-trained GAN model [2, 3], and combines with conditional manipulation operation to achieve more precise control of individual facial attributes, namely *gender, age, smile, pose* and *eyeglasses*, and can even be applied for fixing some generation artifacts of the output image. Moreover, the suggested framework is well-suited for real image editing by leveraging some famous GAN inversion methods [5, 8] and encoder-decoder generative models (e.g. [7]).

By referring to a previous observation that linear interpolation in latent space matches smooth interpolation in image space [6], and by showing that the probability, that a random sample drawn from Normal distribution $N(0, I_d)$ is close enough to the given hyperplane, increases with increasing the dimension d, we come to a key assumption, that for each binary semantic, there is a decision boundary in the (high enough dimensional) latent space, separating between positive and negative samples. We further verify this assumption by training individual *linear SVMs* for each of the desired attributes, and by validating their performance. We also expect that for given semantic, the semantic score $f(\cdot)$ of a generated image $g(z)$ depends linearly on the signed distance of z from the corresponding hyperplane given by its normal n, i.e. $f(g(z)) = \lambda d(n, z)$ (for $\lambda > 0$). Based on this assumption, we show how to perform single attribute manipulation, as well as conditional manipulation in the latent space for image editing. For single semantic, we simply vary the latent code z along the corresponding hyperplane normal direction n as $z_{edit} = z + \alpha n$, thus getting enhanced (decreased) semantic, if $\alpha > 0$ ($\alpha < 0$). Following this further, we also consider the issue of entangled semantics, since editing one attribute might affect the other, e.g. person tends to look older when adding glasses. To address this, we suggest conditional manipulation approach to find a new direction of manipulation along which editing the desired attribute keeps the conditional attributes unaffected. Such a new direction can be found by subtracting from the primal direction its projection onto the plane constructed by all conditional directions (Fig.1 shows the case of one conditional attribute). To sum up, first we synthesize large number of images *(500K)* with randomly sampled latent codes, and assign semantic labels to these images with pre-trained predictor networks. After that, we train linear SVMs in latent space for each of the desired facial attributes by using the assigned labels as ground truth.

Finally, we conduct extensive experiments to evaluate InterFaceGAN framework for two advanced GAN models, namely PGGAN and StyleGAN. First, we make sure that all linear boundaries of both PGGAN Z space and StyleGAN W space show over 95% (75%) accuracy on validation set (entire set), which verifies the assumed separability property of latent space. Then the manipulation ability of the encoded semantics is verified by experiments, both for single attribute (Fig.2) and conditional manipulation case (Fig.3). Notably, we observe an interesting distance effect while moving latent code too far from the boundary, i.e. the original face identity is no longer preserved. The framework is further applied for artifact correction (Fig.4), after learning separation boundary between "good" and "bad" generations (manually labeled 4K "bad" samples). Furthermore, we successfully apply the framework for real image editing by combining different models and inversion methods, and notice that StyleGAN with optimization-based inversion method shows the best results. We thus conclude the superior generalization ability of InterFaceGAN w.r.t. previous approaches.

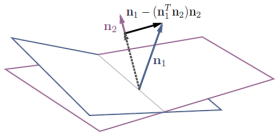

Figure 1: Illustration of conditional manipulation operation for single condition. The new manipulation direction, orthogonal to conditional direction n_2, is found by subtracting from primal direction n_1 its projection onto n_2, i.e. $n_1 - (n_1^T n_2)n_2$.

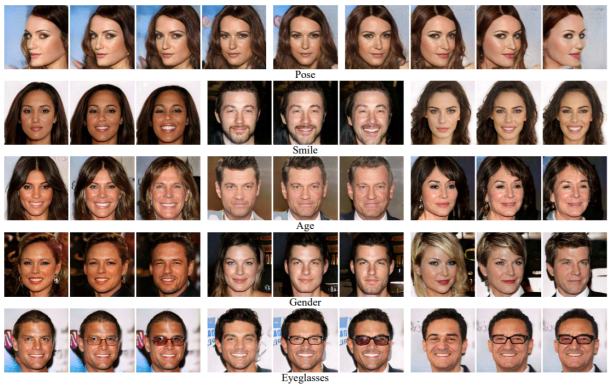

Figure 2: Single attribute manipulation example. The first row illustrates the same face with varied pose in both directions. The other rows depict results of manipulation with four different attributes for different faces, where the central ones are the original generations.

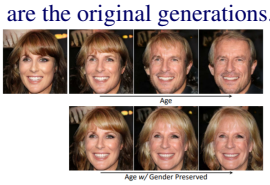

Figure 3: The first row shows manipulation results along original "Age" direction, and the second row shows conditional manipulation of "Age" attribute by preserving "Gender" attribute.

Figure 4: Example of artifact correction by varying latent codes along the better "synthesis quality" direction, learnt by corresponding SVM.

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

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

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

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

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

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

[7] Jiapeng Zhu, Deli Zhao, and Bo Zhang. Lia: Latently invertible autoencoder with adversarial learning, 2019. arXiv preprint arXiv:1906.08090.

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