Interpreting the latent space of GANs for Semantic Face Editing

Group Name **-** Titans.
Group Members **-** Daramalla Nishitha(11940320),
Bandela Santaz Sahithi(11940230).

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 latent code sampled from a random distribution to a photorealistic image. Previous work assumes the latent space learned by GANs follows a distributed representation but observes the vector arithmetic phenomenon. In this work, we propose a novel framework, called InterFaceGAN, for semantic face editing by interpreting the latent semantics learned by GANs. We find that the latent code for well-trained generative models, such as PGGAN and StyleGAN, actually learns a disentangled representation after some linear transformations. Based on our analysis, we propose a simple and general technique, called **InterFaceGAN**, for semantic face editing in latent space. We manage to control the pose as well as other facial attributes, such as gender, age, eyeglasses. More importantly, we are able to correct the artifacts made by GANs.

Introduction:

The popularity of sharing selfies and portrait photos online motivates the rapid development of face editing tools. Facial attribute manipulation is especially attractive with the functions of adding/removing face accessories, such as facial hair and eyeglasses, and/or changing intrinsic face properties, such as age and gender. Facial attribute manipulation has attracted great interest, because of the great chance it brings to research and real-world application. Early work focuses on specific attributes of facial hair generation, expression change, beautification/de-beautification, aging, etc. Recently, with the development of deep neural networks, especially generative adversarial networks, several general face attribute manipulation frameworks were proposed. These approaches take facial attribute edit as an unpaired learning task, and thus are capable of handling different attributes by only changing the data. Our method can be categorized into this group, which aims to provide a general solution for different facial attributes.

Problem Definition:

Given an image(containing face) we need to edit the facial attributes like age, gender, smile etc. by interpreting the latent space of GANs.

Objective:

The main objective of our project is to generate the images and edit their facial attributes. We have to train five independent linear SVMs on pose, smile, age, gender, eyeglasses, and then evaluate them on the validation set (6K samples with high confidence level on attribute scores) as well as the entire set (480K random samples). We also try to visualize some samples by ranking them with the distance to the decision boundary.

Technology used:

- python 3.7
- pytorch 1.1.0
- tensorflow 1.12.2
- sklearn 0.21.2
- Google collab
- Google drive
- Github

Problems faced:

- As the given models are already pre-trained and were saved as pytorch files, we didn't face many problems for collecting the data.
- The GitHub Repository and the Readme that we used for reference is badly documented and hence we faced issues in understanding the code and in it's implementation.
- The readme file didn't mention anything about how to train the model so it was very tough for us to understand how to train the model.

Datasets:

1.CelebA-HQ

The **CelebA-HQ** dataset is a high-quality version of CelebA that consists of 30,000 images with 1024×1024 resolution. To meaningfully demonstrate our results at high output resolutions, we need a sufficiently varied high-quality dataset. However, virtually all publicly available datasets previously used in GAN literature are limited to relatively low resolutions ranging from 322 to 4802. To this end, we created a high-quality version of the CELEBA dataset consisting of 30000 of the images at 1024 × 1024 resolution.

1024 × 1024 images generated using the CELEBA-HQ dataset.

2.FFHQ-(Flickr-Faced-HQ)

Flickr-Faces-HQ (FFHQ) is a high-quality image dataset of human faces, originally created as a benchmark for generative adversarial networks (GAN). FFHQ consists of 70,000 high-quality PNG images at 1024×1024 resolution and contains considerable variation in terms of age, ethnicity and image background. It also has good coverage of accessories such as eyeglasses, sunglasses, hats, etc. The images were crawled from Flickr, thus inheriting all the biases of that website, and automatically aligned and cropped using dlib. Only images under permissive licenses were collected. Various automatic filters were used to prune the set, and finally Amazon Mechanical Turk was used to remove the occasional statues, paintings, or photos.

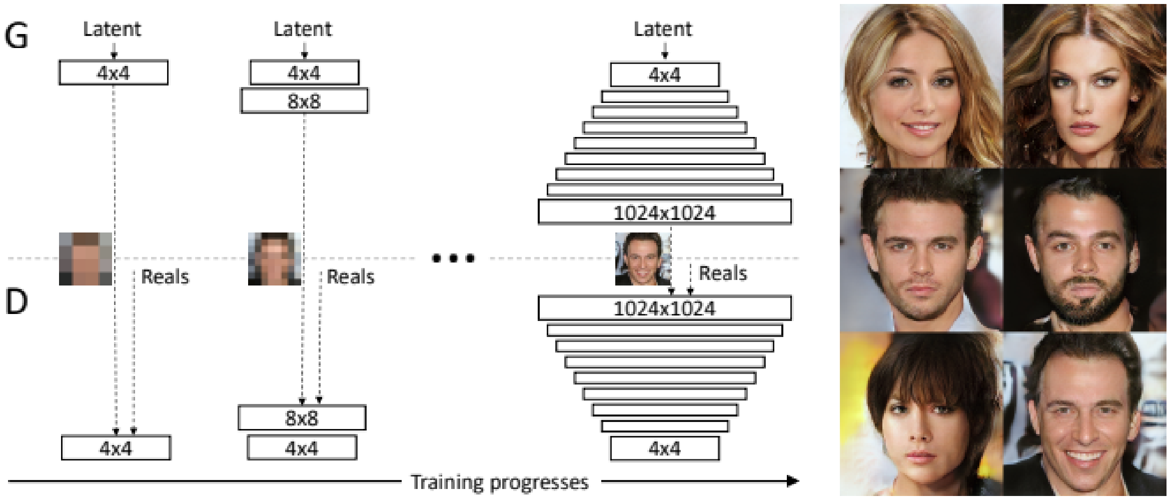
Models:

GAN**-** Generative Adversarial Networks
The generator model in the GAN architecture takes a point from the latent space as input and generates a new image. It basically maps the latent codes (commonly sampled from high-dimensional latent space, such as standard normal distribution) to photo-realistic images.

- Progressive GAN

We describe a new training methodology for generative adversarial networks. The key idea is to grow both the generator and discriminator progressively: starting from a low resolution, we add new layers that model increasingly fine details as training progresses. This both speeds the training up and greatly stabilizes it, allowing us to produce images of unprecedented quality, e.g., CelebA images at 1024^2. We also propose a simple way to increase the variation in generated images, and achieve a record inception score of 8.80 in unsupervised CIFAR10.

Additionally, we describe several implementation details that are important for discouraging unhealthy competition between the generator and discriminator. Finally, we suggest a new metric for evaluating GAN results, both in terms of image quality and variation. As an additional contribution, we construct a higher-quality version of the CelebA dataset.



Training progress in pggan

Our training starts with both the generator (G) and discriminator (D) having a low spatial resolution of 4×4 pixels. As the training advances, we incrementally add layers to G and D, thus increasing the spatial resolution of the generated images. All existing layers remain trainable throughout the process. Here N × N refers to convolutional layers operating on N × N spatial resolution. This allows stable synthesis in high resolutions and also speeds up training considerably. One the right we show six example images generated using progressive growing at 1024 × 1024.

- Style GAN (A Style-Based Generator Architecture for Generative Adversarial Networks)

We propose an alternative generator architecture for generative adversarial networks, borrowing from style transfer literature. The new architecture leads to an automatically learned, unsupervised separation of high-level attributes (e.g., pose and identity when trained on human faces) and stochastic variation in the generated images (e.g., freckles, hair), and it enables intuitive, scale-specific control of the synthesis. The new generator improves the state-of-the-art in terms of traditional distribution quality metrics, leads to demonstrably better interpolation properties, and also better disentangles the latent factors of variation. To quantify interpolation quality and disentanglement, we propose two new, automated methods that are applicable to any generator architecture. Finally, we introduce a new, highly varied and high-quality dataset of human faces.

Different from conventional GANs, StyleGAN proposed a style-based generator. Basically, StyleGAN learns to map the latent code from space Z to another high dimensional space W before feeding it into the generator. Latent space W shows much stronger disentanglement property than Z, since W is not restricted to any certain distribution and can better model the underlying character of real data. We did a similar analysis on both Z and W spaces of

StyleGAN as did to PGGAN and found that W space indeed learns a more disentangled representation. Such disentanglement helps W space achieve strong superiority over Z space for attribute editing.

Implementation:

We choose five key facial attributes for analysis, including pose, smile (expression), age, gender, and eyeglasses. The corresponding positive directions are defined as turning right, laughing, getting old, changing to male, and wearing eyeglasses.To better predict these attributes from synthesized images, we train an auxiliary attribute prediction model using the annotations from the CelebA dataset with the ResNet50 network. This model is trained with multi-task losses to simultaneously predict smile, age, gender, eyeglasses, as well as the 5-point facial landmarks. Here, the facial landmarks will be used to compute yaw pose, which is also treated as a binary attribute (left or right) in further analysis. Besides the landmarks, all other attributes are learned as bi-classification problem with softmax cross entropy loss, while landmarks are optimized with l2 regression loss. As images produced by PGGAN and StyleGAN are with 1024×1024 resolution, we resize them to 256×256 before feeding them to the attribute model.

Given the pre-trained GAN model, we synthesize 500K images by randomly sampling the latent space. There are mainly two reasons in preparing such large-scale data:
(i) to eliminate the randomness caused by sampling and make sure the distribution of the latent codes is as expected
(ii) to get enough wearing-glasses samples, which are really rare in PGGAN model.

To find the semantic boundaries in the latent space, we use the pre-trained attribute prediction model to assign attribute scores for all 500K synthesized images. For each attribute, we sort the corresponding scores, and choose 10K samples with highest scores and 10K with lowest ones as candidates. The reason in doing so is that the prediction model is not absolutely accurate and may produce wrong predictions for ambiguous samples, e.g., middle-aged person for age attribute. We then randomly choose 70% samples from the candidates as the training set to learn a linear SVM, resulting in a decision boundary. Recall that, normal directions of all boundaries are normalized to unit vectors.Remaining 30% are used for verifying how the linear classifier behaves. Here, for SVM training, the inputs are the 512d latent codes, while the binary labels are assigned by the auxiliary attribute prediction model.

Result and Performance:

We trained five independent linear SVMs on pose, smile, age, gender, eyeglasses, and then evaluated them on the validation set (6K samples with high confidence level on attribute scores) as well as the entire set (480K random samples). We can see the results in Table 1. We find that all linear boundaries achieve over 95% accuracy on the validation set and over 75% on the entire set, suggesting that for a binary attribute, there exists a linear hyperplane in the latent space that can well separate the data into two groups.

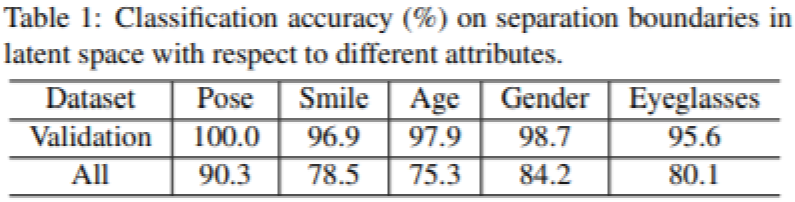
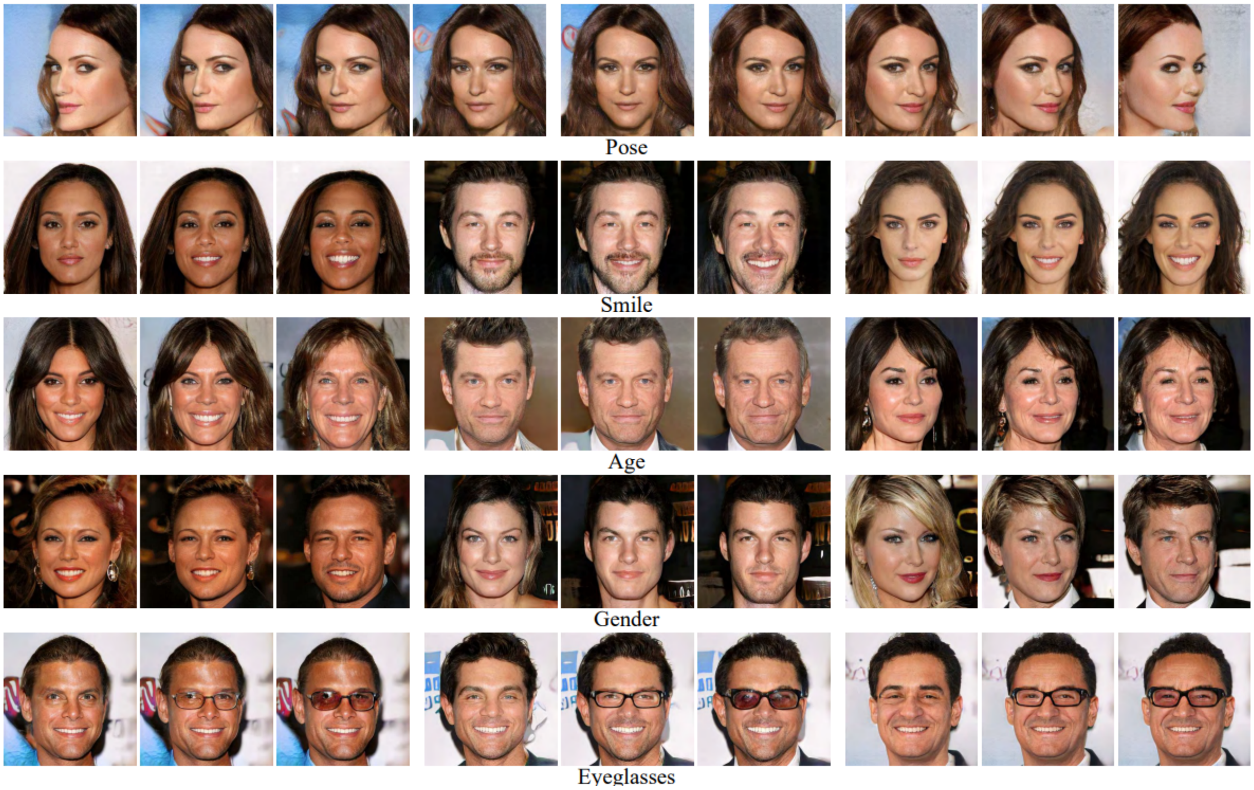


Table 1: Classification accuracy (%) on separation boundaries in latent space with respect to different attributes.

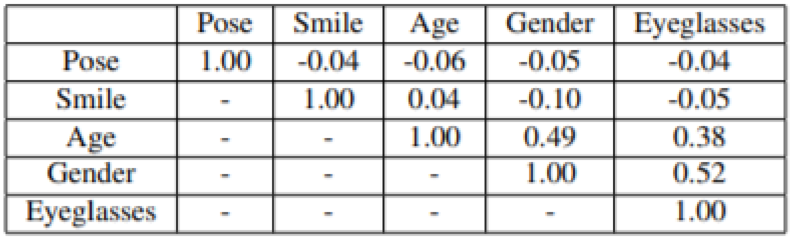
Dataset	Pose	Smile	Age	Gender	Eyeglasses
Validation	100.0	96.9	97.9	98.7	95.6
All	90.3	78.5	75.3	84.2	80.1



In the above figure these are the Single attribute manipulation results. The first row shows the same person under gradually changed poses. The following rows correspond to the results of manipulating four different attributes. For each set of three samples in a row, the central one is the original synthesis, while the left and right stand for the results by moving the latent code along negative and positive direction respectively.

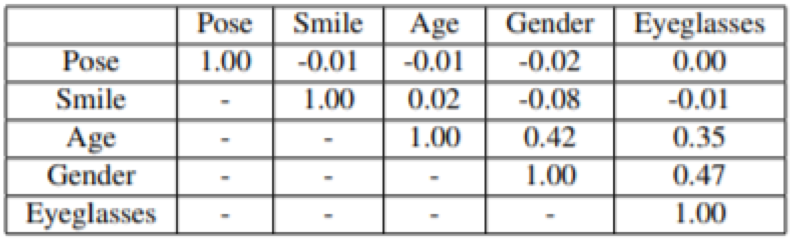
We can tell that attributes behaved similarly under the two metrics, showing that our InterFaceGAN is able to accurately identify the semantics hidden in latent space. We also find that pose and smile are almost orthogonal to other attributes. Nevertheless, gender, age, and eyeglasses are highly correlated with each other. This observation reflects the attribute correlation in the training dataset (i.e., CelebA-HQ) to some extent, where male old people are more likely to wear eyeglasses. This characteristic is also captured by GAN when learning to produce real observation

Table 2: Correlation matrix of attribute boundaries.



	Pose	Smile	Age	Gender	Eyeglasses
Pose	1.00	-0.04	-0.06	-0.05	-0.04
Smile	-	1.00	0.04	-0.10	-0.05
Age	-	-	1.00	0.49	0.38
Gender	-	-	-	1.00	0.52
Eyeglasses	-	-	-	-	1.00

Table 3: Correlation matrix of synthesized attribute distributions.



	Pose	Smile	Age	Gender	Eyeglasses
Pose	1.00	-0.01	-0.01	-0.02	0.00
Smile	-	1.00	0.02	-0.08	-0.01
Age	-	-	1.00	0.42	0.35
Gender	-	-	-	1.00	0.47
Eyeglasses	-	-	-	-	1.00

Conclusions and Further work:

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. Extensive experiments suggest that InterFaceGAN can also be applied to real image editing.

Further work :
We need to implement new facial attribute and check how it works for better understanding.
We also need to check the accuracy of our model.

References:

1. https://genforce.github.io/interfacegan/

2. https://arxiv.org/pdf/1907.10786.pdf

3. https://arxiv.org/pdf/2005.09635.pdf

4. https://github.com/genforce/interfacegan

5. https://github.com/suvojit-0x55aa/celebA-HQ-dataset-download

6.https://paperswithcode.com/dataset/ffhq