Comparison of GANs' implementation on Face Aging

Yihang Hu, Kaiwen Lan, Zenan Wang, Hanji Sun

Instructor: Professor. Li

Abstract

Generative adversarial network (GAN) methods create highly realistic new data from input images. One of its applications in the image-to-image transformation area is the face aging. In the face aging task, aged face images are synthesized by plugging in images from a person's younger years. Face aging can be useful in multiple areas such as in biometric systems, in forensics for finding missing children, in entertainment, and many more. Currently, several GANs are available for face aging applications and this paper focuses on comparing the four GANs which are Cycle-Consistent Adversarial Network (CycleGAN), Attention-Guided Generative Adversarial Network (AttentionGAN), style-based Generative Adversarial Network (StyleGAN), Lifespan Age Transformation Synthesis (LATS). For comparison, these models are trained on the FFHQ (Flickr Faces HQ) dataset. The results of each model are evaluated quantitatively with two image quality assessment metrics (FID and PSNR), and qualitatively with a perceptual study on synthesized images from the same input image. It has been concluded that overall LATS has the best performance over the other models. The code is available at `https://github.com/Psyduck572/STOR566_Final_Project`

Keywords: Generative adversarial network, CycleGAN, AttentionGAN, StyleGAN, Face Aging

1. Introduction

Since the widespread implementation of deep learning, significant development has been achieved in various tasks such as face detection and face recognition. The application of the automatic face aging method with machine learning can be used to handle large database in fields such as auto-readjustment of electronic-records, avoiding contact in attendance system for offices and college classes, official document renewal, electronic customer-retailer business, finding lost children where the children's biological face appearance changes over the years.

As a result, it becomes necessary to go deeper into the existing face aging methods, so that in the future various face aging problems can be solved by methods with better accuracy and image quality. From these considerations, four face aging methods using CycleGAN, AttentionGAN, StyleGAN, and LATS are compared for face aging application and evaluated.

As these four GANs have attracted tremendous attention in image-to-image translation, they have generated remarkable results. To the best of our knowledge, no comparison is done between these four GANs for face aging tasks using the same dataset. In this paper, experiments are conducted which give the comparison analysis between CycleGAN, AttentionGAN, StyleGAN, and LATS to measure the ability to produce plausible and realistic

face aging images.

The first model (CycleGAN) comprises two generators, two discriminators, and converting an image from one domain to another without the need for paired images dataset. The second is AttentionGAN, which consists of attention masks and content masks multiplied with the generated output in one domain to generate a highly realistic image in another domain.

The main objectives of this paper are:

1. The comparison between the different architectures of CycleGAN, AttentionGAN, StyleGAN, and LATS models.

2. Quantitative evaluation by the FID and PSNR score, on the performance of each GAN, using the FFHQ dataset.

3. Qualitative evaluation of the performance of each GAN by generating a few examples to be evaluated by human inspection, using the same input photo.

4. Discuss the potential directions for future study.

2. Related work

2.1 Cycle-Consistent Adversarial Networks (CycleGAN)

CycleGAN has two sets (G and F) of generators and discriminators and is trained with a cycle-consistency loss which measures the difference between the input (p) and recovered input (F(G(p))). CycleGAN can convert an image from one domain to another without using paired images dataset.[1]

2.2 Attention-Guided Generative Adversarial Networks (AttentionGAN)

AttentionGAN identifies the foreground objects and minimizes the changes in the background. Using attention masks, content masks, and the generated output in one domain, AttentionGAN generates highly realistic images in another domain.[2]

2.3 Pivotal Tuning for Latent-based Editing of Real Images(StyleGAN)

It is still challenging to apply ID-preserving facial latent-space editing to faces which are out of the generator's domain, given StyleGan's latent space induces an inherent tradeoff between distortion and editability. Daniel et al. bridged this gap through Pivotal tuning, which preserves the editing quality of an in-domain latent region, while changing its portrayed identity and appearance. In Pivotal Tuning Inversion (PTI), an initial inverted latent code serves as a pivot, around which the generator is fine-tuned. A regularization term is added in the optimization step to keep nearby identities intact. This technique is proven to work well in harder cases, including heavy make-up or elaborate hairstyles.[3]

2.4 Analyzing and Improving the Image Quality of StyleGAN(StyleGAN2)

StyleGANv2 is, as the name suggests, an enhancement of StyleGAN. The images generated with StyleGAN will have droplet artifacts, which sometimes already exist on the feature maps. So the generator of StyleGANv2 removes the AdaIN in the StyleGAN and divides it into two parts: normalization and modulation. Then StyleGANv2 added weight

demodulation to replace normalization, which not only removes artifacts but also retains the full controllability of the image details. Other improvements that make generated images better: Lazy regularization to reduce the amount of calculation, No Progressive growth to make the images more natural, and Path length regularization making it easier to control the properties of the generated images.[4]

3. Dataset

Flickr-Faces-HQ (FFHQ) was used for training our models. The FFHQ dataset consists of 70,000 high-quality PNG images at 1024×1024 resolution and includes various age, ethnicity, image background, and various accessories like eyeglasses, caps etcetera.

4. Proposed Methods

In this section, this project will introduce the four different kinds of generative adversarial networks, maining talking about their architectures. Besides, optimizers and activate functions will also be introduced here.

4.1 CycleGAN

The main obvious difference between regular GAN and the CycleGAN introduced here is that CycleGAN has two generator models and two discriminator models. One generator takes images from the first domain as input and outputs images for the second domain, the other generator takes images from the second domain as input and generates images for the first domain. The purpose of these two discriminators is to determine how plausible the generated images are and update the generator models accordingly. Furthermore, one more advanced technique applied in CycleGAN is cycle consistency, which is used to guide the image generation process in the new domain toward image translation. Hence, as the existence of this property, the CycleGAN does not need paired images to do the whole process, but instead, unpaired images can also be useful.

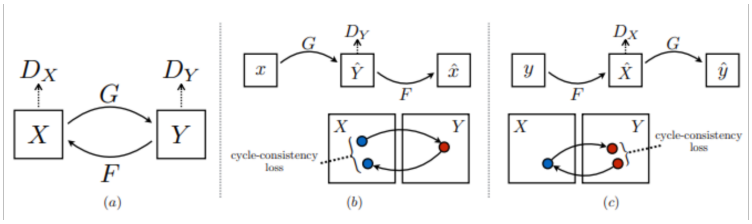


Figure 1: CycleGAN Architecture[1]

The figure above shows the whole process of applying CycleGAN. The image in Domain X will be used as an input, and then the Generator G will use the input to train and generate an output, placing it in Domain Y. After that, the Discriminator in Domain Y will test the output. If it successfully fools the Discriminator in Domain Y, it will again be used as an input in Generator F to train and generate the final generated image, and put it

back in Domain X. At last, the cycle-consistency loss will check if the final generated image is the same as the original image. The above process is just one direction. The CycleGAN will implement the method in both directions like what the figure (b) and (c) show below. Hence, that is the whole cycle process of the CycleGAN.

4.1.1 Activation Function

In the project's CycleGAN, it applies the Leaky ReLU as its activation function instead of ReLU. Leaky ReLU has a small slope for negative values, instead of altogether zero, which regular ReLU has. It fixes the "dying ReLU" problem, which occurs when learning rate is too high or there is a large negative bias, since it does not have zero-slope parts. Besides, it also speeds up the training. The reason is that unlike ReLU, leaky ReLU is more "balanced," and may therefore learn faster.

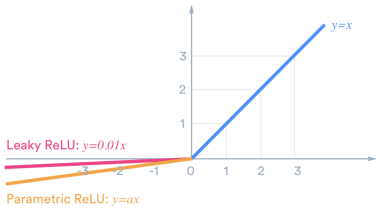


Figure 2: Leaky ReLU Activation function[7]

From the figure above, it shows that unlike ReLU, Leaky ReLU has a small slope for negative values, instead of altogether zero, which can improve the "dying ReLU" problem and also reduce the training time.

4.1.2 Optimizer

Here, the CycleGAN uses the Adam optimization algorithm, which is a classical stochastic gradient descent as its optimizer. Since Adam optimizer has a faster computation time, and requires fewer parameters for tuning, it has the best probability of getting the best results for the proposed model.

4.2 AttentionGAN

The main novelty of the AttentionGAN is the implementation of attention masks based on the general framework of CycleGAN. In AttentionGAN, two mappings between domains X and Y via two generators are learned , i.e., $G : x \rightarrow [A_y, C_y] \rightarrow G(x)$ and $F : y \rightarrow [A_x, C_x] \rightarrow F(y)$, where A_x and A_y are the attention masks of images x and y, respectively; C_x and C_y are the content masks of images x and y, respectively; $G(x)$ and $F(y)$ are the generated images, like what it shows in figure 3.

The attention masks A_x and A_y define a per pixel intensity specifying to which extent each pixel for the content masks C_x and C_y will contribute to the final output image.

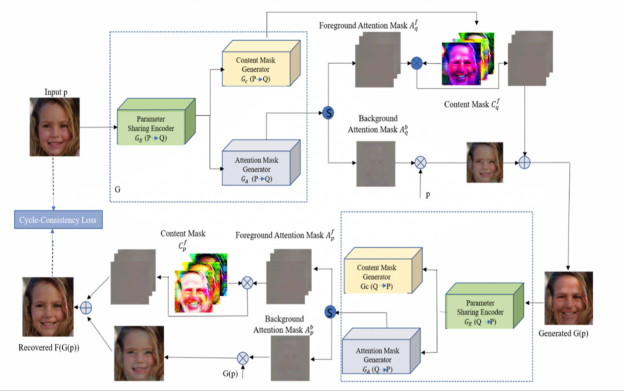


Figure 3: AttentionGAN Architecture[6]

Consequently, the generator won't change static objects (which is the background) and can focus on the pixels defining the domain content movements, leading to more accurate and more realistic output images.

Generating both foreground and background attention masks, the model changes the foreground while preserving the background of a given face image. Mathematically, content and attention masks are multiplied with the generated image from the generator to synthesize a realistic image, the formula is the following equation:

$$G(p) = \sum_{f=1}^{n-1}(C_q^f * A_q^f) + p * A_q^b$$

Where C is the content mask, A is the attention mask, and p is the input image. The activation function and optimizer of AttentionGAN are similar to that of CycleGAN as it is based on the architecture of CycleGAN.[2]

4.3 StyleGAN utilizing Pivotal Tuning

4.3.1 Pivotal tuning

Performing strictly in the generator's latent space, GAN inversion results in distortion. Pivotal Tuning is a novel approach: a pretrained StyleGAN is tuned , such that the input image is generated when using the pivot latent code found in the previous off-the-shelf inversion techniques. The pivotal tuning is a local operation in the latent space, shifting the identity of the pivotal region to the desired one with minimal compensation.

The figure above shows that StyleGAN's latent space is portrayed in two dimensions, where the warmer colors indicate higher densities of W, i.e. regions of higher editability. On the left, we can see the Editability-Distortion trade-off. A choice must be made between Identity "A" and Identity "B". "A" resides in a more editable region but does not resemble the "Real" image. "B" is the opposite. On the right, "C" maintains the same high editing capabilities of "A", while achieving even better similarity to "Real" compared to "B".

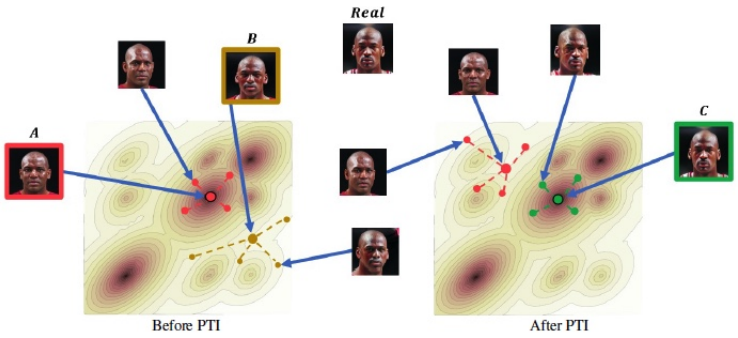


Figure 4: Pivotal Tuning Illustrator[3]

4.3.2 Method

First, we invert the given input to wp in the native latent space of StyleGAN, W. The purpose of the inversion step is to provide a convenient starting point for the Pivotal Tuning, since StyleGan's native latent space W provides the best editability. To tune the StyleGAN model that is pre-trained on FFHQ dataset, a direct optimization is applied to optimize both latent code w and noise vector n to reconstruct the input image x, measured by the LPIPS perceptual loss function.

$$w_p, n = \underset{w,n}{\text{argmin}} \mathcal{L}_{LPIPS}(x, G(w, n; \theta)) + \lambda_n \mathcal{L}_n(n)$$

where G(w, n, θ) is the generated image using a generator G with weights theta.[3]

Second, applying the latent code w obtained in the inversion, produces an image that is similar to the original one x, but may yet exhibit significant distortion. Therefore, in the second step, we unfreeze the generator and tune it to reconstruct the input image x given the latent code w obtained in the first step, which we refer to as the pivot code wp. Let $x^p = G(w_p; \theta^*)$ be the generated image using w_p and the tuned weights[3] . We fine tune the generator using the following loss term:

$$\mathcal{L}_{pt} = \mathcal{L}_{LPIPS}(x, x^p) + \lambda_{L2}\mathcal{L}_{L2}(x, x^p)$$

4.3.3 Locality Regularization

The visual quality of images generated by non-local latent codes, which are used in Pivotal Tuning, is compromised. A regularization term is introduced, designed to restrict the PTI changes to a local region in the latent space.

4.4 Lifespan Age Transformation Synthesis (LATS)

Roy et al.[5]developed a new GAN-based method, multi-domain image-to-image conditional GAN, in Lifespan Age Transformation Synthesis designed to simulate the process of continuous aging from a single input image. This model, trained with FFHQ-aging, was designed to fantasize about the aging process and generate an approximate appearance

throughout a person's life cycle. The project used the model developed by Roy et al. as a pre-trained model in the experimental phase to compare with other GANs.

Principle–The dataset used in this model is partitioned into 10 age intervals (age groups), and when a target age is given, a vector age code is assigned to each image of each age group and sent to the mapping network, which is then sent to the latent space of the target age to achieve a continuous age transformation. The generator of the model is used to transform the age groups and consists of an identity encoder and a decoder. The identity encoder extracts the features related to the person's identity, while the decoder uses the same network structure as StyleGAN2 [4] to inject identity into the output images and also to avoid droplet artifacts.

5. Experiments

In this section, the project puts the preprocessed dataset in several models, including CycleGAN, AttentionGAN, StyleGAN, and LATS, to compare their performance to get the best result.

5.1 Experiment design

EXPERIMENT 1—After training CycleGAN and AttentionGAN on the same dataset, the project applies Frechet Inception Distance (FID) and Peak Signal-to-Noise Ratio (PSNR) to evaluate the results.

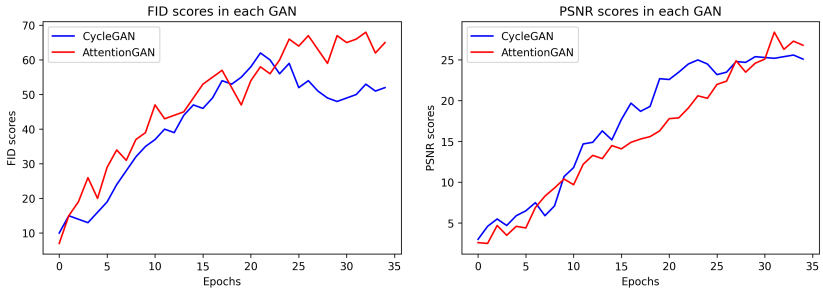
EXPERIMENT 2—The quantitative comparison analysis may not be straightforward. Besides, since the StyleGAN and LATS the project used here are pretrained models, the project does the visualization. It uses some sample images to do the face aging using these four models and print out the outputs to do the comparison, checking if they are plausible or not.

5.2 Evaluation Metric

For image quality assessment, first of all, the project uses Frechet Inception Distance (FID) to do the evaluation. It tells how real the generated face images are in comparison to real face images. FID is obtained by computing a "distance" between the generated image and the real image. The lower value represents the better quality of the image. Furthermore, Peak Signal-to-Noise Ratio (PSNR) and Mean-Square Error (MSE) is introduced here to evaluate our results. The MSE represents the cumulative squared error between the compressed and the original image, whereas PSNR represents a measure of the peak error. The lower the value of MSE, the lower the error. Here is the formula:

$$PSNR = 10 \times log_{10}(\frac{max^2}{MSE})$$

Besides, it will print out some images we transformed and evaluate them by ourselves to see if the aging process is successful or not.



	CycleGAN	AttentionGAN
FID	50.8	67.3
PSNR	25.22	26.82

5.3 Results and Comparison

The Frechet Inception Distance (FID) score is got by computing a "distance" between the generated image and the real image, which means the smaller the score value the better. But the higher the PSNR value the better, the higher the PSNR represents the less distortion after compression.

In this project, CycleGAN and AttentionGAN were trained for 35 epochs, and the performance was almost the same, but the actual data showed that CycleGAN was slightly better than AttentionGAN in generating face-aging images.

The picture on the right is a comparison of 4 GANs in generating face-aging images. Obviously, both StyleGAN and LATS perform significantly better than CycleGAN and AttentionGAN. Especially LATS, which is better than StyleGAN in the processing of hair texture details, and it is also considered to be the best of these 4 GANs. The image generated by CycleGAN has missing pixels, while the image generated by AttentionGAN is blurred. As for the droplet artifact in LATS mentioned in the presentation, it may be caused by the preprocessing of LATS. The preprocessing of LATS is to remove the background and clothes of the training images.

6. Conclusions

In general, CycleGAN outperforms AttentionGan in terms of resolution and quality; LATS (StyleGANv2) demonstrates stable performance through the removal of water droplet artifacts and better preservation of hairstyle and skin texture, compared to StyleGAN. To guarantee better performance, gender and ethnicity will be taken into account: the model will be pre-trained on certain datasets that only contain pictures of only one race of a fixed gender to allow the model to distinguish between races and genders.

7. References

[1] Welander P, Karlsson S and Eklund A 2018 Generative Adversarial Networks for Image-to-Image Translation on Multi-Contrast MR Images - A Comparison of CycleGAN and UNIT. arXiv:1806.07777

[2] H. Tang, H. Liu, D. Xu, P. H. S. Torr and N. Sebe, "AttentionGAN: Unpaired Image-to-Image Translation Using Attention-Guided Generative Adversarial Networks," in IEEE Transactions on Neural Networks and Learning Systems, doi: 10.1109/TNNLS.2021.3105725.

[3] Daniel Roich, Ron Mokady, Amit H. Bermano, and Daniel Cohen-Or. 2022. Pivotal Tuning for Latent-based Editing of Real Images. ACM Trans. Graph. 42, 1, Article 6 (February 2023), 13 pages. https://doi.org/10.1145/3544777

[4] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, Timo Aila; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8110-8119

[5] Or-El, R., Sengupta, S., Fried, O., Shechtman, E., Kemelmacher-Shlizerman, I. (2020). Lifespan Age Transformation Synthesis. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, JM. (eds) Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science(), vol 12351. Springer, Cham. https://doi.org/10.1007/978-3-030-58539-6$_4$4

[6] Sharma, N., Sharma, R. Jindal, N. Comparative analysis of CycleGAN and AttentionGAN on face aging application. Sādhanā 47, 33 (2022). https://doi.org/10.1007/s12046-022-01807-4

[7] Danqing Liu, A Practical Guide to ReLU, published: Nov 30, 2017, Medium