Facial transformation with CycleGAN

Zhengye Zhu Peking University
Guixian Chen Peking University
Jiayu Su Peking University
Weiyi Yan Peking University

Abstract

In our project, we use a popular method (CycleGAN) to change specific characteristics of one face such as age, gender and race. CycleGAN is an awosome algorithm that can learn to translate images between domains, although in the absence of any paired training examples. We improve the traditional model to transfer between three or more domians. What's more, we want to apply face detection to real-world images to change many faces in batch. It can be used not only for entertainment, but also for cross-age face recognition, criminal investigation, etc.

1 Introduction

It's awesome knowing how we'll look when we're getting old. That's why many people just can't get enough of the FaceApp old age filter. Inspired by this, we want to change faces using deep learning methods. We regard this task as Image-to-image translation problem, which is a class of vision and graphics problems where the goal is to learn the mapping between an input image and an output image.

At present, there are many ideas about facial transformation, among which, GAN is the most popular one. Without doubt, CycleGAN is the most mainstream approach [1], so we choose it as our major research object. Besides, RsGAN [11] is used in face swapping and editing in R.Natsume,T.Yatagawa and S.Moirshima's work. DiscoGAN is also widely concerned for it is able to learn the relevance of unsupervised data from different domain and its similarity with CycleGAN.

We apply CycleGAN to transform images in two different domains like young and old. In order to transfer between more domains, we adjusted its structure and propagation process. Finally, we get models that can change facial features such as age, gender, race, etc.

2 Related Work

2.1 Face aging

Face aging, also known as age progression, aims to render a given face image with natural aging effects under a certain age or age group. In recent years, face aging has attracted major attention due to its extensive use in numerous applications, entertainment, finding missing children, cross-age face recognition, etc. Although impressive results have been achieved recently, there are still many challenges due to the intrinsic complexity of aging in nature and the insufficient labeled aging data. Antipov et al. [2] introduce a novel approach for "Identity-Preserving" optimization of GAN's latent vectors. The objective evaluation of the resulting aged and rejuvenated face images by the state-of-the-art face recognition and age estimation solutions demonstrate the high potential of the proposed method. Zhang et al. [3] proposes a conditional adversarial autoencoder (CAAE) that learns a face manifold, traversing on which smooth age progression and regression can be realized simultaneously

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

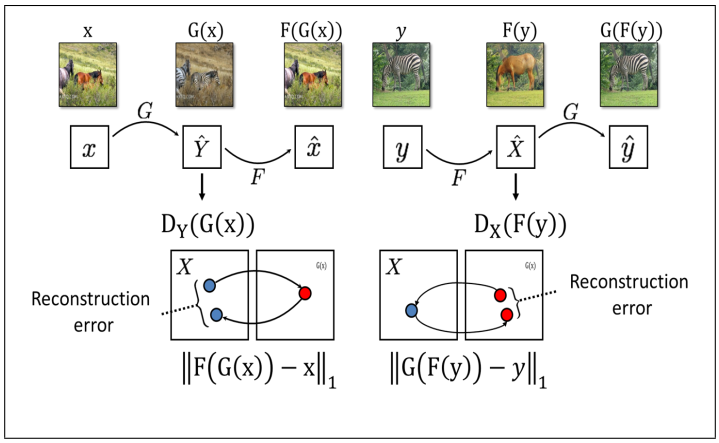

Figure 1: CycleGAN

2.2 Generative Adversarial Network

GANs [4] have achieved impressive results in image generation, image editing, and representation learning. Recent methods adopt the same idea for conditional image generation applications, such as text2image, image inpainting, and future prediction, as well as to other domains like videos and 3D data. The key to GANs' success is the idea of an adversarial loss that forces the generated images to be, in principle, indistinguishable from real photos. This loss is particularly powerful for image generation tasks, as this is exactly the objective that much of computer graphics aims to optimize.

2.3 Image-to-Image Translation

The idea of image-to-image translation goes back at least to Hertzmann et al.'s Image Analogies [5], who employ a non-parametric texture model on a single input-output training image pair. More recent approaches use a dataset of input-output examples to learn a parametric translation function using CNNs. CycleGAN approach builds on the "pix2pix" framework of Isola et al. [6], which uses a conditional generative adversarial network to learn a mapping from input to output images.

2.4 CycleGAN

For many Image-to-image translation tasks, paired training data will not be available. CycleGAN [1] is an approach for learning to translate an image from a source domain X to a target domain Y in the absence of paired examples. As shown in Fig 1, it's goal is to learn a mapping $G : X \rightarrow Y$ such that the distribution of images from G(X) is indistinguishable from the distribution Y using an adversarial loss. Because this mapping is highly under-constrained, they couple it with an inverse mapping $F : Y \rightarrow X$ and introduce a cycle consistency loss to enforce $F(G(X)) \approx X$ (and vice versa).

2.5 Face detection

Face detection is a computer technology being used in a variety of applications that identifies human faces in digital images. Face detection can be regarded as a specific case of object-class detection. In object-class detection, the task is to find the locations and sizes of all objects in an image that belong to a given class. R-CNN [7] apply high-capacity convolutional neural networks to bottom-up region proposals in order to localize and segment objects. YOLO [8] uses a single neural network predicts bounding boxes and class probabilities directly from full images in one evaluation.

3 Formulation

3.1 Adversarial loss

We apply adversarial losses to both mapping functions. For the mapping function $G : X \rightarrow Y$ and its discriminator D_Y , we express the objective as:

$$L_{GAN}(G, D_Y, X, Y) = E_{y \sim p_{data}(y)}[logD_Y(y)] + E_{x \sim p_{data}(x)}[log(1 - D_Y(G(x))],$$

where G tries to generate images $G(x)$ that look similar to images from domain Y , while D_Y aims to distinguish between translated samples $G(x)$ and real samples y. G aims to minimize this objective against an adversary D that tries to maximize it, i.e., $min_G max_{D_Y} L_{GAN}(G, D_Y, X, Y)$. We introduce a similar adversarial loss for the mapping function $F : Y \rightarrow X$ and its discriminator D_X as well: i.e., $min_F max_{D_X} L_{GAN}(F, D_X, Y, X)$.

3.2 Cycle Consistency loss

As shown in Fig 1, for each image x from domain X, the image translation cycle should be able to bring x back to the original image, i.e., $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$. It's forward cycle consistency. We incentivize this behavior using a cycle consistency loss:

$$L_{cyc}(G, F) = E_{x \sim p_{data}(x)}[||F(G(x)) - x||_1] + E_{y \sim p_{data}(y)}[||G(F(y)) - y||_1].$$

3.3 Identity loss

We find that it is helpful to introduce an additional loss to encourage the mapping to preserve color composition between the input and output. We regularize the generator to be near an identity mapping when real samples of the target domain are provided as the input to the generator:

$$L_{identity}(G, F) = E_{y \sim p_{data}(y)}[||G(y) - y||_1] + E_{x \sim p_{data}(x)}[||F(x) - x||_1].$$

3.4 Loss

Our full objective is:

$$L(G, F, D_X, D_Y) = L_{GAN}(G, D_Y, X, Y) + L_{GAN}(F, D_X, Y, X) + \lambda_1 L_{cyc}(G, F) + \lambda_2 L_{identity}(G, F),$$

where λ_i controls the relative importance of the losses. We aim to solve:

$$G^*, F^* = arg \min_{G,F} \max_{D_x, D_Y} L(G, F, D_X, D_Y).$$

4 Implementation

4.1 Models

As shown in Fig 2, traditional CycleGAN is a two-domain model. For our project, we proposed a cross-domain model, which has more cycles.

Following the architecture of CycleGAN, our network contains two stride-2 convolutions, several residual blocks , and two fractionallystrided convolutions with stride 1/2. We use 6 blocks for 128 $\times$ 128 images and 9 blocks for 256 $\times$ 256 and higherresolution training images. Similar to Johnson et al. [9], we use instance normalization. For the discriminator networks we use 70 $\times$ 70 PatchGANs [22, 30, 29], which aim to classify whether 70 $\times$ 70 overlapping image patches are real or fake. Such

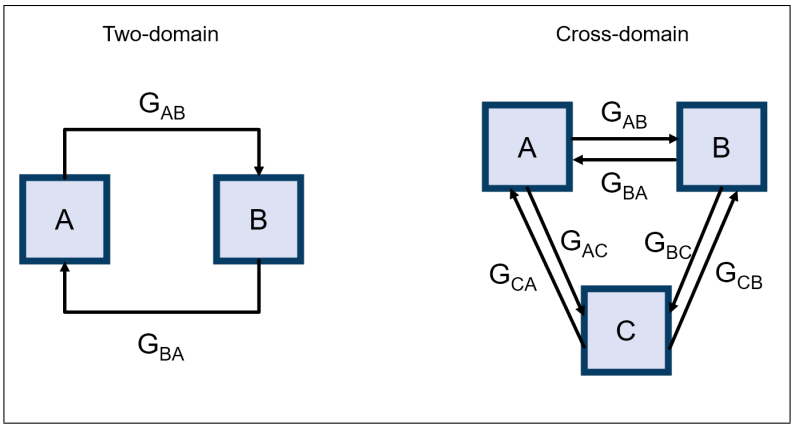

Figure 2: models

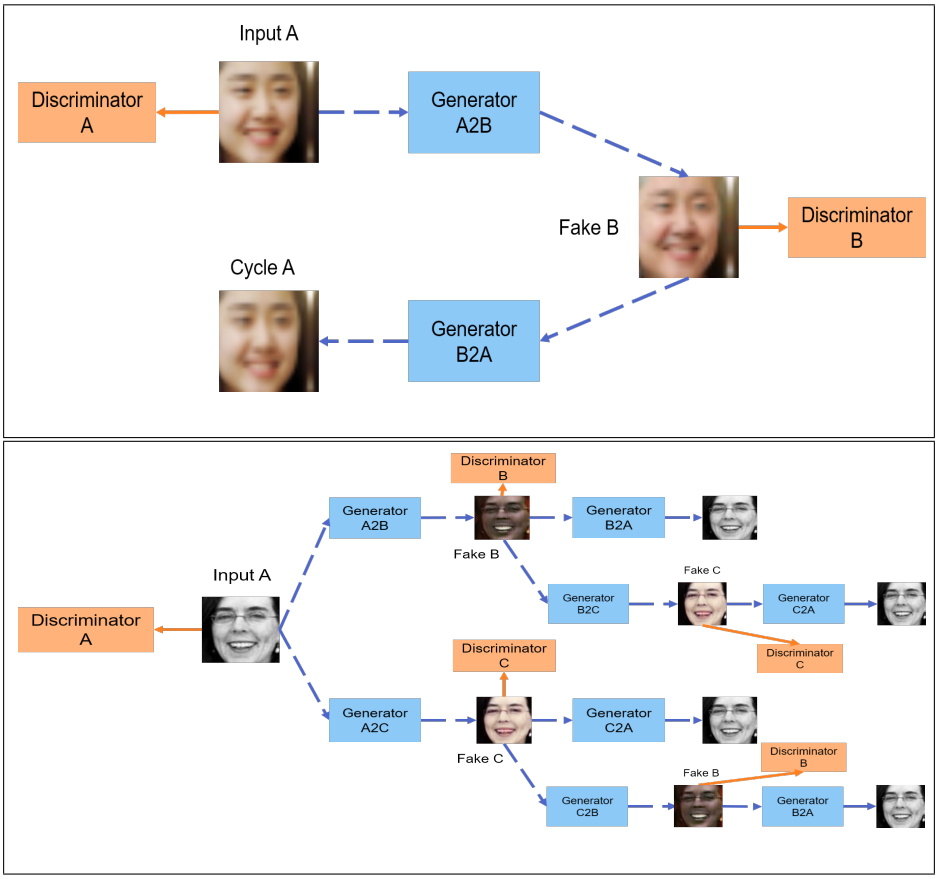

Figure 3: propagation process

a patch-level discriminator architecture has fewer parameters than a full-image discriminator and can work on arbitrarily-sized images in a fully convolutional fashion [6].

4.2 Propagation

As shown in Fig 3, when inputting an image from A domain into two-domain models, we will get a fake image from B domain. Further, we can then get an image from A domain, which is assumed to be the same as the input image. With these images, we can easily calculate the losses. As to the cross-domain model, when inputting an image from A domain, we will get fake images from both B

Figure 4: Face Detection: There are some traces of bboxes.

and C. Further, we can then get CycleA images. In order to utilize Generators between B and C, we generated new fake images based on fakeB and fakeC images. Although we only take input A as an example, the process of inputting B or C is the same.

4.3 Dataset

We train our models on UTKFace: Large Scale Face Dataset [3]. UTKFace dataset is a large-scale face dataset with long age span (range from 0 to 116 years old). The dataset consists of over 20,000 face images with annotations of age, gender, and ethnicity. The images cover large variation in pose, facial expression, illumination, occlusion, resolution, etc. This dataset could be used on a variety of tasks, e.g., face detection, age estimation, age progression/regression, landmark localization, etc. For our project, we split the dataset into young and old, male and female, white, yellow and black, etc.

4.4 Training

Following the CycleGAN, for L_{GAN}, we replace the negative log likelihood objective by a least-squares loss. This loss is more stable during training and generates higher quality results. To reduce model oscillation, we follow Shrivastava et al.'s strategy [10] and update the discriminators using a history of generated images rather than the ones produced by the latest generators.

For all the experiments, we set $\lambda_1 = \lambda_2 = 10$. We use the Adam solver with a learning rate of 0.0002 and a batch size of 1(because of the instance normalization).

4.5 Face detection

In real-world it might not always be possible to get such images to use our CycleGAN for face-changing. We need to be able to find where a face is present in an image and modify that part of the image. For this we will run a face detector before passing the image to CycleGAN. The face detector gives bounding boxes of the various faces in an image. We take crops of those boxes to send it to our network. We then take the outputs to place it back on the input image. This way we can deal with any image from real world. We use opencv face-detector [12] which is based on resnet-ssd architecture.

5 Results

Due to hardware limitations, we only trained less than 100 epochs. From fig 5, we can see losses of generators are much larger than of discriminators. As time goes on, the loss is decreasing more and more slowly.

Due to unpaired images, there are not efficient methods to evaluate our results. However, we can roughly evaluate by our eyes and it is just meaningful for practical applications. Fig 6 shows some of the results. As we can see, the converted images have a certain fidelity, while retaining the original features well.

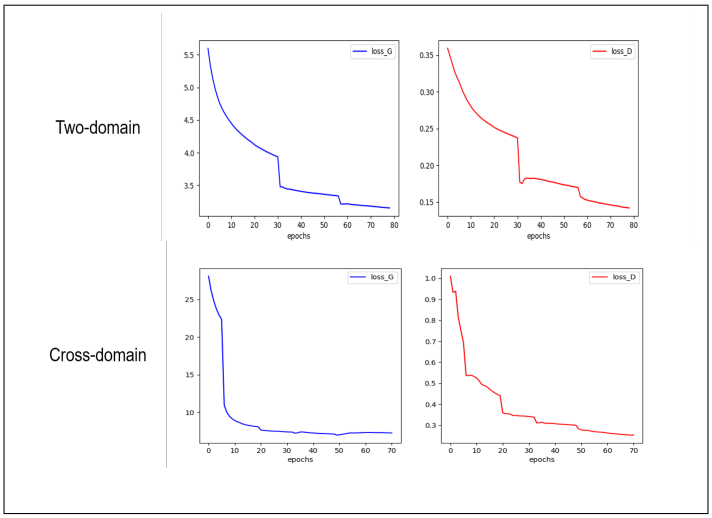

Figure 5: losses

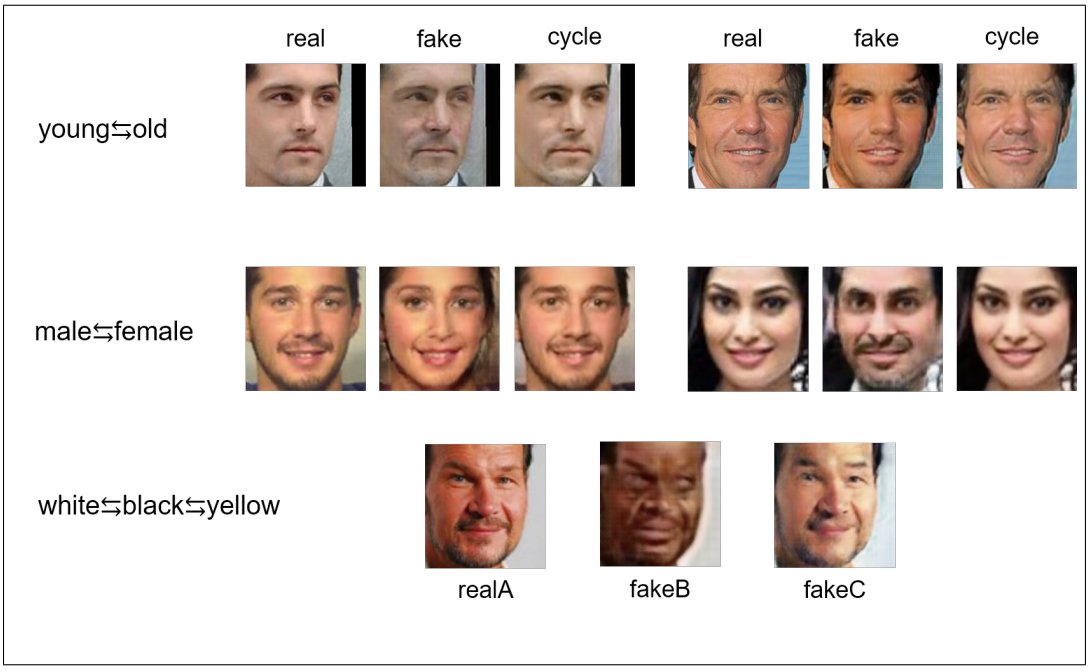

Figure 6: results

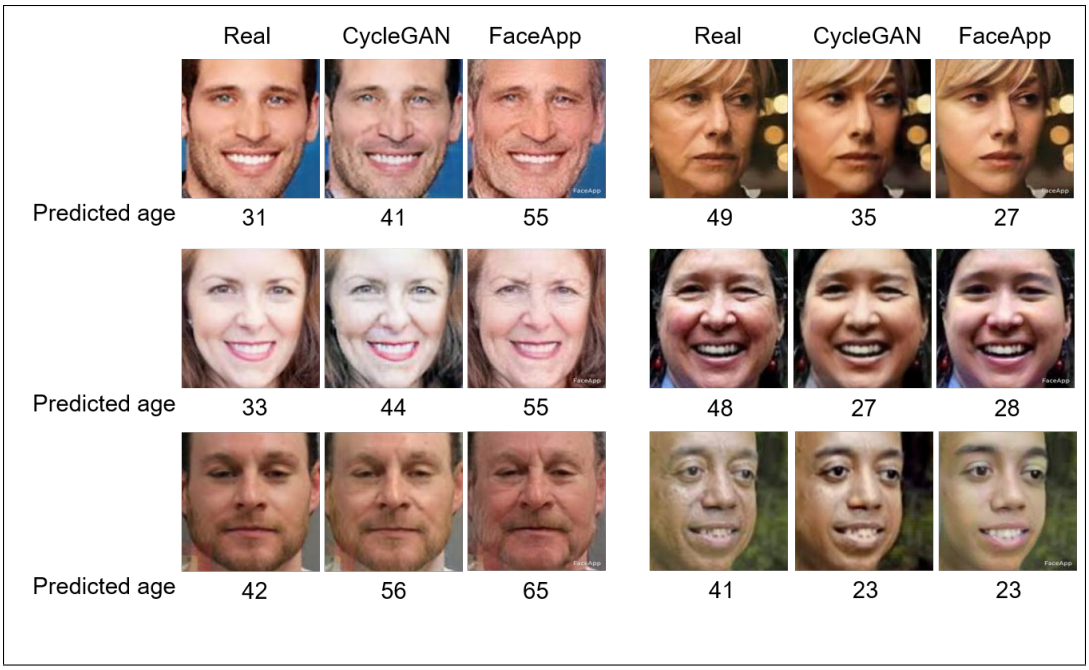

Figure 7: results

To evaluate the performance of our model, we compared the results of CycleGAN with images transformed by the commercialized facial transformation tool FaceApp (FaceApp Age-Young and Age-Old filters). In particular, we applied Microsoft's How-Old age predictor to quantitatively assess the quality of facial transformations. As shown in Fig 7, qualitatively, our model tends to be much softer, and the facial contours remain basically unchanged when adding or removing age features. On the contrary, FaceApp's results seem to be more extreme (older or younger), and some corrections are made to the contours. Both methods slightly adjust the character's skin color, while in our model, the change is somehow larger. More quantitative comparisons of facial transformations using the How-Old age predictor confirmed our preliminary observation. Generally, the age of the pictures generated by our model is closer to the age of the original pictures, while the age of FaceApp transformation results tends to be far away.

6 Improvements

There may be several possible improvements:

- adopt better training strategys, like Wasserstein GAN
- adjust weights of different losses
- try super-resolution image generation
- look for better quantitative evaluation indicators

7 Conclusion

In our project, we build a model to implements the facial transformation function by using CycleGAN, which successfully complete the task of image-to-image translation problem. As for application, our model is capable of age conversion, gender transition, skin color conversion and so on. It can be applied to entertainment, such as sharing the modified picture on social software. And it can be applied to cross-age face recognition, which is more accurate and practical than traditional model. Moreover, it can be used in criminal investigation, photo inspection and other areas. All in all, our model has satisfactory accuracy, convenience, and practicality.

8 Author Contributions

We conceived of the presented idea together. Chen investigated multiple papers and theories. Zhu wrote the models and trained them on colab. Yan and Su tested and evaluated the models. We also discussed the results and solutions to some bugs together. All authors contributed to the presentation and final report.

References

[1] Jun-Yan Zhu, Taesung Park, Phillip Isola & Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on, 2017*.

[2] Antipov G, Baccouche M & Dugelay J L. Face aging with conditional generative adversarial networks[C] *2017 IEEE international conference on image processing (ICIP)*. IEEE, 2017: 2089-2093.

[3] Z. Zhang, Y. Song & H. Qi, "Age Progression/Regression by Conditional Adversarial Autoencoder," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp. 4352-4360, doi: 10.1109/CVPR.2017.463.

[4] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C] *Advances in neural information processing systems*. 2014: 2672-2680.

[5] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless & D. H. Salesin. Image analogies. In SIGGRAPH, 2001.

[6] P. Isola, J.-Y. Zhu, T. Zhou & A. A. Efros. Imageto-image translation with conditional adversarial networks. In CVPR, 2017.

[7] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 580-587).

[8] Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788).

[9] Johnson J , Alahi A , Fei-Fei L . Perceptual Losses for Real-Time Style Transfer and Super-Resolution[J]. 2016.

[10] Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., & Webb, R. (2017). Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2107-2116).

[11] Natsume R , Yatagawa T , Morishima S . RSGAN: Face Swapping and Editing using Face and Hair Representation in Latent Spaces[J]. 2018.

[12] Bradski, G. (2000). The OpenCV Library. Dr. Dobb's *Journal of Software Tools*.