Transforming Facial Images Using Generative Adversarial Networks

Teng Ma, Qingyuan Jin, Isha Srivastava, Jiaqi Yan
University of California, San Diego
Group 34
{tema, q1jin, isrivast, j3yan}@eng.ucsd.edu

Abstract

Image-to-image translation has recently become extremely popular in various fields like research, entertainment and photography assistance. In our project, we have implemented one of the most popular image-transferring methodologies in machine learning- the cycle-GANs transferring technology. In this report, we talk about the conceptual idea of building the model, how to implement the methodology, some exhibits of training results and test results, and our further discussion upon experiment with hands-on experience. To see more details, visit our project at GitHub .

1. Introduction

Many problems in image processing, computer graphics, and computer vision can be posed as translating an input image into a corresponding output image. Just as a concept may be expressed in either English or French, a scene may be rendered as an RGB image, a gradient field, an edge map or a semantic label map. [1]. Along similar lines, image-to-image translation can be defined as a class of vision and graphics problem, where the goal is to learn a mapping that can convert an image from a source domain to a target domain while preserving the main characteristics of the input image. Some of the popular applications of image-to-image translation include image stylization, image segmentation, season/style transfer, photo enhancement and object transfiguration.

One particular application that we will be concentrating on, in this paper, is face transfer. Face transfer is an extension of image-to-image translation, where the goal is to transform facial images from one domain to another. More specifically, it can be described as generating new facial images with realistic characteristics from original faces that look at least superficially authentic to human observers. In our project, the two said domains are those of males and females. Thus, our task for this project can be formally defined as, given a domain X of male images, generate new images that would closely resemble female images in domain Y and vice versa. Figure 1 shows an example of face transfer between females and males.

Figure 1. **Example of face transfer from female to male.**

For our project, we have used Generative Adversarial Networks (GANs) which given a training set, learn to generate new data with the same statistics as the training set. The specific model that we have used is CycleGAN which is explained in detail in Section 4.

Our input to this model (CycleGAN), was male images from domain X and our output was closely resembling female images from domain Y. Similarly, we also input female images from domain Y into the CycleGAN model to get new male images.

2. Related Work

In this section, we discuss prior work that is closely related to our problem at hand. We have organized our Literature review into 4 subsections in which we discuss all the concepts that we have used in an orderly manner.

2.1. Unsupervised Image-to-Image translation

A classic approach to unsupervised representation learning is to do something like K-mean clustering on the data. In the context of images, hierarchical clustering of image patches as shown in [2] can be used to learn powerful image representations. In [3], the authors pointed out that unsupervised image-to-image translation aims at learning a joint distribution of images in different domains by using images from the marginal distributions in individual domains. To address the problem of an infinite set of joint distributions possibly arriving at the given marginal distributions,

they proposed an unsupervised image-to-image translation framework based on Coupled GANs.
Like all the above work, in this paper, we also formulate face transfer as an unsupervised image-to-image translation task.

2.2. Face Transfer in videos

In [4], the authors used multi-linear model of 3D faces to perform face transfer in videos. They were able to achieve this by tracking a face template and re-rendering it under different expression parameters. Later, the authors of [5] investigated expression modeling and face trackers to transfer facial expressions and achieved real-time face transfer.
All the above works are related to face transfer. However, one fundamental difference between their and our work is that our problem focuses on face transfer in images rather than videos. Also, compared with above works, our approach uses CycleGANs, which is based on deep learning, to achieve face transfer in images, without any supervision.

2.3. Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GAN) [6] has recently been getting a lot of attention in unsupervised learning tasks. Conditional GAN, as a variant of GAN, is widely used in various computer vision scenarios. In some image-to-image translation tasks, the inputs are images rather than noises. [7], [10], [11] investigated similar cycle architecture and named this architecture as CycleGAN, DiscoGAN, DualGAN respectively.

In this paper, the model that we have used is CycleGAN as explained in [7].

2.4. Latest Approach - TL-GANs

In our opinion, one of the best approaches for the given task of face transfer is a fairly new one. Introduced in October 2018, it is called Transparent Latent-space GAN (TL-GAN) [12]. This is the state-of-the-art approach and is significantly different from the previous works because it allows users to gradually tune one or multiple features using a 'single network', as opposed to one specific GAN network for one feature. This simplifies the model architecture, as tuning multiple features can now be done simultaneously. Besides, the author claims that adding new tunable features can be done very efficiently in less than one hour. This is one of the approaches that we have left as future work for our project.

3. Dataset and Features

In this project, we have used celebrity images from the CelebFaces Attributes Dataset (CelebA) [13]. This is a large scale face attributes dataset with more than 200K celebrity images. Figure 2 shows few sample images from our dataset.

Figure 2. **Sample images from the CelebA dataset.**

3.1. Split for training and testing

For our task, we have used a total of 1900 images, out of which, there are 747 male celebrity images and 1153 female celebrity images. Out of these, we have removed 100 images of each category (male and female) for testing prior to the training step. Since this is an unsupervised learning task, we do not require a validation set intrinsically and 100 test images of each category are enough to evaluate how well our model generalizes and performs on unseen data.

3.2. Preprocessing

The size of each image from the CelebA dataset is 218 x 178 and contain extra background attributes other than just the face. So, cropping faces from such images would result in smaller, lower-resolution pictures. In order to avoid this problem, we have used HD CelebA Cropper [14] to obtain higher resolution facial images of celebrities. After this preprocessing step, we get square images of the face having a resolution of 512 x 512 pixels. Figure 3 shows few images from our dataset after using HD CelebA Cropper.
Thus, our input data includes colored images that are not normalized having a resolution of 512 x 512 pixels.

3.3. Features

As far as the features are concerned, GANs learn the underlying structure of the given data without specifying a target value. CycleGANs can automatically learn features combining many aspects properly like colors, corners and edges. Therefore, we did not have to extract any features in our project.

Figure 3. **Sample images from the CelebA dataset post preprocessing (HD CelebA cropper).**

4. Method

4.1. Generative Adversarial Networks (GAN)

The key idea to GANs is an adversarial loss that forces the generated images to be 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. We adopt an adversarial loss to learn the mapping such that the translated images cannot be distinguished from images in the target domain [7].

Cycle-GANs' loss functions contain adversarial loss and cycle consistency loss, here we introduce some several types of methods to derive loss function, one of the mostly used is vanilla loss. For mapping function G and its discriminator D_Y, the vanilla loss function is

$$\mathcal{L}_{GAN\_Vanilla}(G,D) = \mathbb{E}_{y\sim p_{data}(y)}[logD(y)] + \mathbb{E}_{x\sim p_{data}(x)}[log(1-D(G(x))] \quad (1)$$

We are aimed at solving the following problem,

$$G^*, D^* = arg\min_G \max_D \mathcal{L}_{GAN}(G,D) \quad (2)$$

For the loss of discriminators,

$$\mathcal{L}_{D\_Vanilla}(D) = -\mathbb{E}_{y\sim p_{data}(y)}[logD(y)] - \mathbb{E}_{x\sim p_{data}(x)}[log(1-D(G(x))] \quad (3)$$

For the loss of generators,

$$\mathcal{L}_{G\_Vanilla}(G) = \mathbb{E}_{x\sim p_{data}(x)}[log(1-D(G(x))] \quad (4)$$

4.2. Deep Convolutional GAN (DCGAN)

DCGAN mainly consists convolution layers without max pooling or fully connected layers. It uses transposed convolution and convolution stride for the down sampling and the up sampling. DCGAN eliminates fully connected layers and uses batch normalization except the output layer for the generator and the input layer of the discriminator. In DCGAN, all max poolings are replaced with convolutional strides. [8]

4.3. Least Square GAN (LSGAN)

Regular GANs adopt the sigmoid cross entropy loss function, which will cause the problem of vanishing gradients for the samples that are on the correct side of the decision boundary, but still far from the real data. [9] LSGAN is an impovement to remedy this problem, which lead that LSGANs can generate more realistic images than regular GANs.

Its discriminator loss function can be defined as follows,

$$\mathcal{L}_{D\_LSGAN}(D) = \mathbb{E}_{y\sim p_{data}(y)}[(1-D(y))^2] + \mathbb{E}_{x\sim p_{data}(x)}[D(G(x)] \quad (5)$$

Its generator loss function can be defined as follows,

$$\mathcal{L}_{G\_LSGAN}(G) = \mathbb{E}_{x\sim p_{data}(x)}[(D(G(x))-1)^2] \quad (6)$$

Two main benefits of LSGANs are listed here. First, LSGANs will penalize those correctly classified samples. In LSGANs, the parameters of the discriminator are fixed when updating the generator. As a result, the penalization will make the generator to generate samples towards the decision boundary. Secondly, moving the generated samples towards the decision boundary results in making them be closer to the manifold of real data [9].

4.4. Cycle-GANs

For cycle-GANs, the loss function combines both discriminator and generator loss,

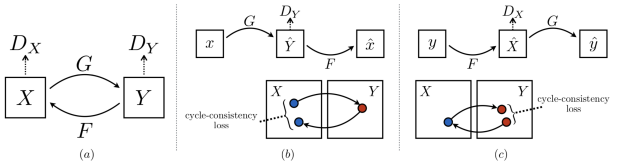


Figure 4. **Cycle-GANs model**

Given training samples $\{x_i\}_{i=1}^N$ where $x_i \in X$ and $\{y_i\}_{i=1}^N$ where $y_i \in Y$, we want to learn mapping functions between domain X and domain Y. Let the two mapping functions be $G : X \mapsto Y$ and $F : Y \mapsto X$. Also, create adversarial discriminators D_X to distinguish between

images x and translated images $F(y)$, and D_Y to distinguish between images y and translated images $F(x)$ [7].

Generally, the loss of cycle-GANs can be described as,

$$\mathcal{L}_{GAN\_cycGAN}(G,F) = \mathcal{L}_{G\_GAN}(G) + \mathcal{L}_{G\_GAN}(F) \tag{7}$$

The cycle consistency loss $\mathcal{L}_{cyc}$ is to make sure that when we translate a image form one domain to the another and then translate back to original domain, the reconstructed images should return the same image like the input picture which means hold the same information in this translation procedure.

$$\begin{aligned}\mathcal{L}_{cyc}(G,F) = \mathbb{E}_{x\sim p_{data}(x)}[\|F(G(x)) - x\|_1] \\ +\mathbb{E}_{y\sim p_{data}(y)}[\|G(F(y)) - y\|_1]\end{aligned} \tag{8}$$

The identity loss $\mathcal{L}_{identity}$ is used to measure the difference between input images and output images without gender transformation. For example, we input a male facial image into female-to-male generator and get a generated male image. We want the generator recognize the gender of input data and do not change the photo.

$$\begin{aligned}\mathcal{L}_{identity}(G,F) = \mathbb{E}_{x\sim p_{data}(x)}[\|F(x) - x\|_1] \\ +\mathbb{E}_{y\sim p_{data}(y)}[\|G(y) - y\|_1]\end{aligned} \tag{9}$$

For overall generator loss of cycleGAN can be discribed as

$$\begin{aligned}\mathcal{L}_{G\_cycGAN} = \mathcal{L}_{GAN\_cycGAN} \\ + \lambda_1\mathcal{L}_{cyc} + \lambda_2\mathcal{L}_{identity}\end{aligned} \tag{10}$$

For overall discriminator loss of cycleGAN can be discribed as

$$\mathcal{L}_{D\_cycGAN}(G,F) = \mathcal{L}_{D\_GAN}(D_X) + \mathcal{L}_{D\_GAN}(D_Y) \tag{11}$$

5. Experiments

There are several experiments on vanilla CycleGAN, least-square CycleGAN, CycleGAN with U-Net Generator and WGAN-GP CycleGAN in our papar.

For all experiments, we set λ_1 = 10 and λ_2 = 0.5 in Equation 10. Adam optimizer is chosen to be used as optimizer of generator and discriminator due to the fact that the ability of Adam optimizer to tweak the steps based on the different delta back-warded from the last layer of the models. The initial learning rate of both models is 0.0002.

Each experiment holds the same training procedure with two major steps to train our model. Firstly, train the generator networks by minimizing GAN loss $\mathcal{L}_{GAN\_cycGAN}$, the cycle loss $\mathcal{L}_{cyc}$ and the identity loss $\mathcal{L}_{identity}$ to make the generators generate fake images which can fool the discriminators and keep the original information except the gender. L_{cyc} and $\mathcal{L}_{identity}$ are same for all models and $\mathcal{L}_{GAN\_cycGAN}$ depends on the basic GAN model of each experiment. Secondly, train the discriminator networks by minimizing $\mathcal{L}_{D\_cycGAN}(G,F)$ to strengthen discriminator and find the fake images.

5.1. Vanilla CycleGAN

In this model, we use Vanilla GAN as our basic GAN model to optimize our parameters. The architecture for our generative networks is the same as Jun-Yan et al. [7] which adopt from Johnson et al. [15] and have great results for style transfer and super-resolution. This network contains three strided Convolutional layers, nine residual blocks and three fractionally strided Convolutional layers. And we use instance normalization just like Jun-Yan et al. [7] and Johnson et al. [15]. For discriminator network, we use the 70×70 PatchGANS [16] which have fewer parameters compared with 1×1 PixelGAN.

5.2. least-square CycleGAN

We use the same architecture as same as Vanilla CycleGAN and using LSGAN loss to train single GAN network.

5.3. CycleGAN with U-Net Generator

In this model, the previous generator is replaced with the U-Net generator. U-Net [17] is a convolutional autoencoder with skip connections. The encoder will downsample the image to a samll size at the bottleneck and then the decoder will upsample the latent feature maps to original size. Every decoder layer has a skip connection with one layer of encoder which have the same size of feature map. And in this experiment, we also use LSGAN to train each basic GAN.

6. Results and Discussion

The results of transforming facial images between male and female facial images are demonstrated in Figure 5 and 6, which can clearly show the difference between three different implemented networks. As shown in Figure 5 and 6, the first column of male facial images is the input image and the following columns are the generated female images of LS-cycleGAN, Vanilla-cycleGAN and cycleGAN based on U-Net generator.

Firstly, we can clearly see the relationship between the original images and correspondent fake images. The most information exclude the gender have been kept in transformation. However, we can see the huge difference between three models. In the last two rows of Figure 5, these have the best performance in this picture. The details of fake images from LS-cycleGAN is much plentiful than the other models. For example, the original skin color has been kept in the fake image. However, in the other results, the skin color

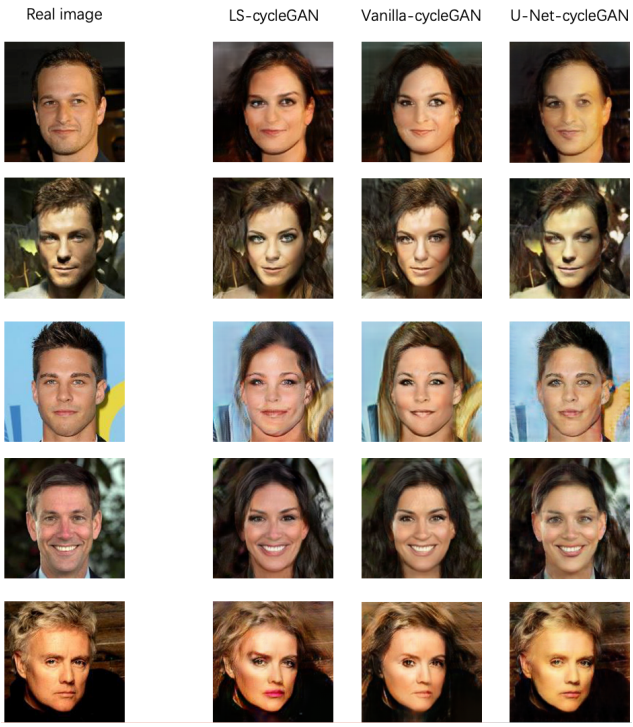


Figure 5. **Demonstration of results for transferring male face to female face.**

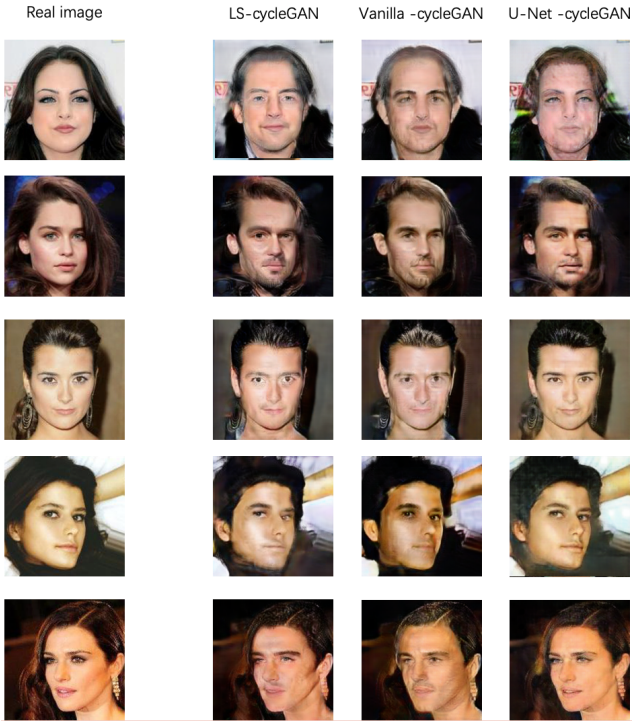
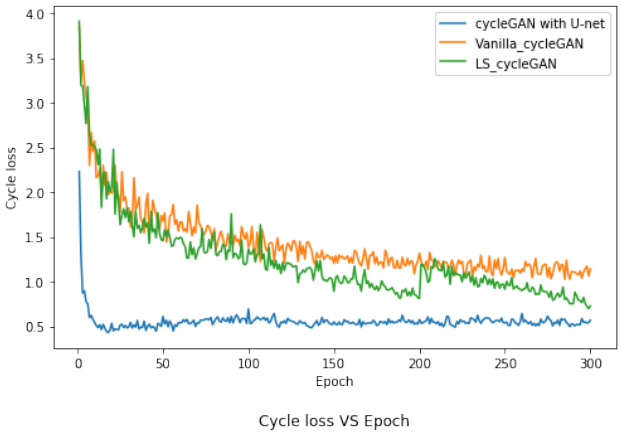


Figure 6. **Demonstration of results for transferring female face to male face.**

become lighter compared with the original image. Using Vanilla-cycleGAN, we didn't get the best result. The possible reason is the Vanilla-GAN loss is the cross entropy loss which focus on the correct classification rate. As long as the discriminator classified the fake images correctly, the loss and gradient will be small and the model will not be trained more detailed. The last column is the results of a totally different generator. Although U-net have much more parameters than residual generator, the output only have slightly change in face. It cannot make a huge transform of hair.

Generally, we can see the female generated images from male face are a little better than the male generated images from our observation. This is because model is hard to transfer hair and results in a male face with female hair which make the picture look strange. We can conclude that for all models, reducing or increasing hair is the hardest step to do gender transformation. The reason for this problem is that the hair information is much complicated than the facial information. The difference of hairstyle in female images is much complex than that of male images. Even female can handle male hairstyle easily. But it is quite weird for men with woman's hair.

And the results on test data is not satisfied as train data. We think this is due to our limited amount of train data which is not sufficient to represent the whole face collection. So the model cannot learn enough knowledge to fit different type of faces.



Cycle loss VS Epoch

Figure 7. $\mathcal{L}_{cyc}$ **of three models**

Figure 7 shows the cycle loss of three models. We can see that the cycle loss of U-net cycleGAN model decreasing rapidly to the lowest point. Maybe this is the reason why U-net cycleGAN model have the worst performance among these models. The cycle loss predominates the total loss which make the output images hold too many information more than we want.

Figure 8, 9, 10 and 11 show the generator loss and discriminator loss of three different models. The most different feature of GAN model compared with other model is

the loss does not converge. To train generator will minimize the generator loss. However, the discriminator Loss will increase because the generator become stronger to fool the discriminator. The loss cannot show the training status like other models.

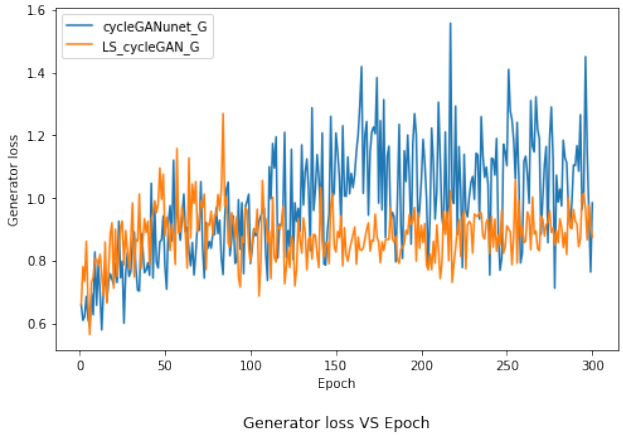


Figure 8. **Generator Loss of LS-cycleGAN and U-net cycleGAN**

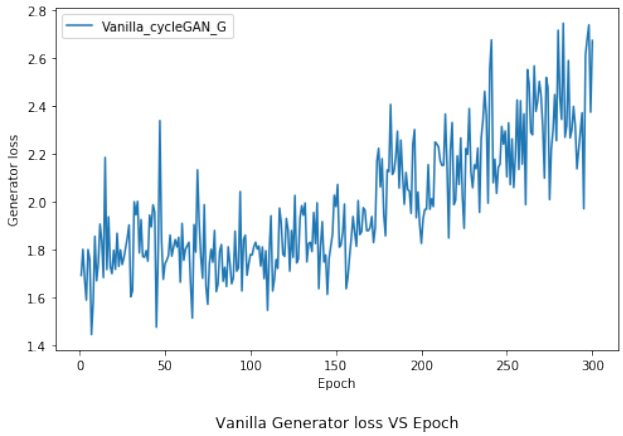


Figure 9. **Generator Loss of Vanilla-cycleGAN**

7. Conclusion

In this paper, we compared three different models to transfer face images between male and female. The LS-cycleGAN with residual generator have the best performance with the most plentiful details.

For future work, we want to find a better optimizer method, apply mini-batch, and scale down the output image to speed up the slow training procedure and tune the hyper-parameter to generate more detailed and realistic images. And explore more edge-cutting algorithms to generate more realistic and stable facial images, such as TL-GAN which can tune one or multiple features using a single network and the newest and most powerful GAN model BigGAN which show really impressive and realistic results.

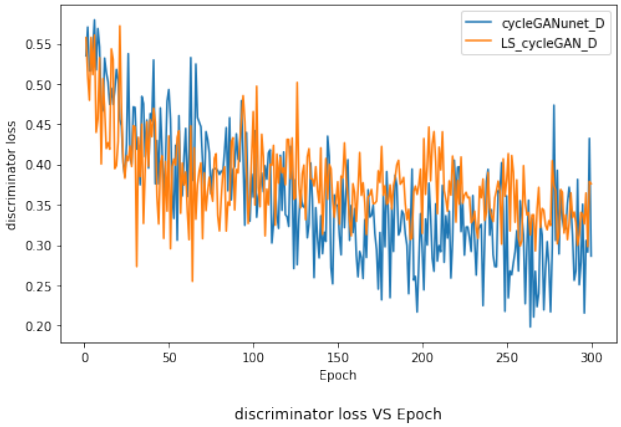


Figure 10. **Discriminator Loss of LS-cycleGAN and U-net cycleGAN**

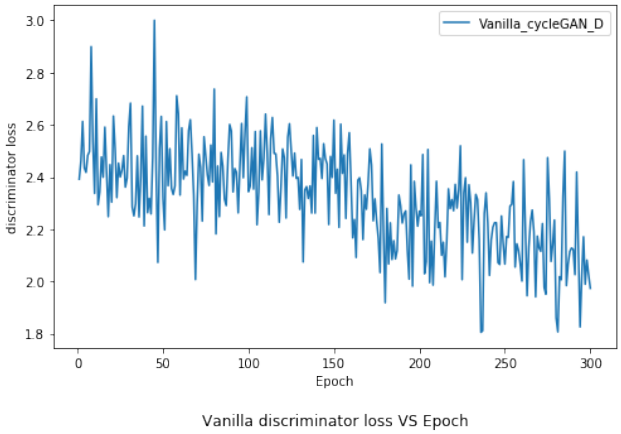


Figure 11. **Discriminator Loss of Vanilla-cycleGAN**

8. Individual Contributions

In this project, Teng finished the LS-cycleGAN amd U-net cycleGAN model. Qingyuan developed Vanilla Model. Isha did data pre-processing and experimental testing to tune model parameters. Jiaqi explored experiment results and did comparison and selection.

9. Acknowledgment

We would like to express our gratitude towards Professor Peter Gerstoft for his patient guidance. We would also like to thank our TAs Ruixian Liu, Siva Prasad Varma Chiluvuri and Harshul Gupta for their kind help.

References

[1] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou and Alexei A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks", arXiv:1611.07004v3, November 2018.

[2] Adam Coates and Andrew Y. Ng, "Learning feature representations with k-means. In Neural Networks: Tricks of the Trade", pp. 561580, Springer, 2012.

[3] Ming-Yu Liu, Thomas Breuel and Jan Kautz, "Unsupervised Image-to-Image Translation Networks", 31st Conference on Neural Information Processing Systems (NIPS 2017), 2017.

[4] Daniel Vlasic, Matthew Brand, Hanspeter Pfister and Jovan Popovic, "Face transfer with multilinear models", ACM transactions on graphics (TOG), volume 24, 426433, ACM, 2005.

[5] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt and Matthias Niebner, "Face2Face: Real-time Face Capture and Reenactment of RGB Videos", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 23872395, 2016.

[6] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville and Yoshua Bengio, "Generative Adversarial Nets", Advances in neural information processing systems, 26722680, 2014.

[7] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks 2017 IEEE International Conference on Computer Vision (ICCV), 2017.

[8] Alec Radford, Luke Metz and Soumith Chintala. "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks".

[9] Xudong Mao1 Qing Li, Haoran Xie, Raymond Y.K. Lau, Zhen Wang, Stephen Paul Smolley, "Least Squares Generative Adversarial Networks", ICCV, 2017

[10] T. Kim, B. Kim, M. Cha and J. Kim, "Unsupervised visual attribute transfer with reconfigurable generative adversarial networks", arXiv preprint arXiv:1707.09798, 2017.

[11] Z. Yi, H. Zhang, P.T. Gong, et al, "Dualgan: Unsupervised dual learning for image-to-image translation", arXiv preprint arXiv:1704.02510, 2017.

[12] Shaobo Guan, Generating custom photo-realistic faces using AI, Medium, 2018.

[13] Ziwei Liu, Ping Luo, Xiaogang Wang and Xiaoou Tang, "Deep Learning Face Attributes in the Wild", Proceedings of International Conference on Computer Vision (ICCV), December 2015.

[14] Zhenliang He, "HD CelebA Cropper", available at https://github.com/LynnHo/HD-CelebA-Cropper

[15] J. Johnson, A. Alahi, and L. Fei-Fei, Perceptual Losses for Real-Time Style Transfer and Super-Resolution, Computer Vision ECCV 2016 Lecture Notes in Computer Science, pp. 694711, 2016.

[16] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, Image-to-Image Translation with Conditional Adversarial Networks, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[17] O. Ronneberger, Invited Talk: U-Net Convolutional Networks for Biomedical Image Segmentation, Informatik aktuell Bildverarbeitung fr die Medizin 2017, pp. 33, 2017.