Rethink Cycle-Consistent Adversarial Networks with Hair Segmentation

Lelei Zhang
Department of Computer Science
Simon Fraser University
leleiz@sfu.ca

Shiying Tu
Department of Computer Science
Simon Fraser University
britneyt@sfu.ca

Abstract

Cycle Consistent Adversarial Networks [1] are widely used to the mappings between two unpaired domains X and Y. However it regards the whole image as an object, therefore loses the transformation information in the detail areas of the domains. In this project, we improve CycleGAN by applying Fully Convolutional Networks for Semantic Segmentation model [2] to guide CycleGAN to focus on the target areas it should transform. Qualitative comparisons with prior models: CycleGAN and DiscoGAN are presented to illustrate our improvements. Also, we get the quantitative results by evaluating our model on CelebA datasets [3].

1 Introduction

1.1 Problem

Hair coloring is one of the most popular cosmetic practices. Unlike other, this practice will still for several months or more. In this case, before people want to try a new hair color, they may be worried if that color is suitable for them. Therefore, we generate the idea to build a model that is able to simulate people appearance after changing their hair color.

With this interest, we select Generative Adversarial Networks (GAN) to do this Image-to-Image translation. There are several powerful GANs that can perform an attribute transformation. However the problem that we lack paired images for the hair transformation arises, which means there are no ground truth images for comparison. Thus, we choose CycleGAN which can learn the mappings between two unpaired domains, to help us build a model that translates between two hair colors: blonde and black.

During training, we found CycleGAN cannot accurately identify the hair region in the image. If the image is someone with very shining hair band, CycleGAN will keep shining part unchanged while dying rest of the hair. One possible reason for the above is that CycleGAN treats the whole picture as an object and transfers the domain implicitly. Therefore, we try to add instance segmentation to CycleGAN to indicate a clearer boundary and then completely transform the attribute we selected.

1.2 Understanding CycleGAN

The power of Cycle-consistent Generative Adversarial Networks lies in being able to learn such transformations without one-to-one mapping between training data in source (domain X) and target domain (domain Y). It contains two mapping functions G: X -> Y and F: Y -> X and associated discriminators D_Y and D_X. The generators perform translation to the opposite domain and try to fool the corresponding discriminator which is trained to classify the given image is real or fake. In our project, we define black hair as domain X and blonde hair as domain Y. By following the architecture presented in Figure 1a), generator G translates black hair (X) to blonde hair (Y) and then we pass the generated fake blonde hair image to discriminator D$_Y$ to determine its validity. Same procedure can also be applied to blonde hair (Y) to black hair (X).

Adversarial losses are applied to both mapping functions. For generator G and discriminator D$_Y$, the loss objective is the following:

$$\mathcal{L}_{\text{GAN}}(\text{G}, D_Y, X, Y) = E_{y\sim p_{data}(y)}[\log D_Y(y)] + E_{x\sim p_{data}(x)}[\log(1 - D_Y(G(x))] \quad (1)$$

First, we want to maximize the accuracy of discriminators for distinguishing the real and fake images. Secondly, the model minimizes the loss for generator since G tries to make generated images G(x) to look as similar as images from domain Y possible. Thus, the final loss is the following:

$$min_G max_{D_Y} \mathcal{L}_{\text{GAN}}(\text{G}, D_Y, X, Y) \quad (2)$$

Similar adversarial loss is introduced for generator F and discriminator D$_X$:

$$min_F max_{D_X} \mathcal{L}_{\text{GAN}}(\text{F}, D_X, Y, X) \quad (3)$$

Cycle Consistency Losses are newly added by CycleGAN to reduce the space of possible mapping functions. According to the large capacity of dataset, the set of images can be mapped to any random permutation of target domains. CycleGAN avoids this by comparing the reconstructed images to their original ones. As illustrated in Figure 1b) and 1c), there is a forward cycle-consistent loss, i.e x → G(x) → $F(G(x)) \approx x$ as well as a backward cycle-consistent loss, i.e y → F(y) → $G(F(y)) \approx y$. Together, we need to minimize the following objective:

$$\mathcal{L}_{\text{cyc}}(\text{G}, F) = E_{x\sim p_{data}(x)}\left[\left\|\text{F}(\text{G}(\text{x})) - x\right\|_1\right] + E_{y\sim p_{data}(y)}\left[\left\|\text{G}(\text{F}(\text{y})) - y\right\|_1\right] \quad (4)$$

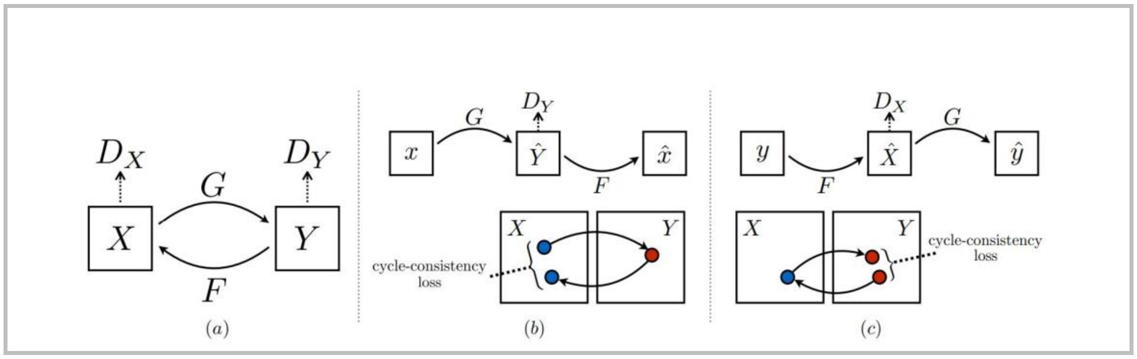

Figure 1: CycleGAN Architecture

1.3 Semantic Segmentation

Different from classification, semantic segmentation links each pixel in an image to a class label. In traditional fully convolutional neural network, image segmentation will take two steps: firstly, features of images are learnt by convolutional neural networks. Then, deconvolution layers are used for upsampling to enable pixelwise prediction.

More efficient convolutional neural networks, MobileNets, is used in our project to condense the hair segmentation. The model is developed by Howard. et al, which replaces the standard convolutional filters by two layers: depth wise convolution and pointwise convolution to build a depth wise separable filter, resulting the effect of a drastic reduction in computation and model size.

2 Approach

In this section, we will describe our contributions in detail. Limitations of CycleGAN will be described and our modifications to it will be explained.

2.1 CycleGAN with sigmoid Discriminators

Original CycleGAN uses 70x70 PatchGAN, which aims to classify whether 70 x 70 overlapping images are real or fake with fewer parameters than the fully convolutional fashion. The detailed architecture is C64-C128-C256-C512. After the last layer, a convolution is added to map to a 1x1 stride output layer. Since we only select the blonde hair and black hair as our two domains. We select the sigmoid activation, which can map the output between 0 and 1, to the output layer.

2.2 CycleGAN with Hair Segmentation

CycleGAN treats the whole given image as a single object and thus it may lose the details in the transformation. For example, as can be seen from first row of Figure 5, it only transfers part of the hair to blonde, while the other part remains black. Also, in last row of Figure 4, when the lady wears a black hat, CycleGAN failed to change the hair color to black because the model considers the black hat as black hair and thus 'skip' this example. Overall, CycleGAN has a vague direction for translation. As a solution, we introduce a semantic segmentation model to guide it focusing on the part we want.

As described in section 1.3, FCN model takes representation learnt by deep convolutional networks and performs pixelwise prediction for given images. Initially we tried to use VGG16 as our encoder, which is a very deep convolutional network for large-scale image with high accuracy in classification (23.7% top-1 val. error) and localization (25.3% top-5 localization error) [4]. Then we found, though with high accuracy, a VGG16 model has around 138 million parameters to evaluate and occupies around 500MB memory, thus not very efficient from this project's perspective. MobileNets, however, turns out to be a very good encoder with only 4.2 million parameters and 70.6% accuracy in ImageNet classification, a little bit lower than VGG16 71.5% ImageNet accuracy [5]. In this project, we use a Fully Convolutional Networks Semantic Segmentation model with MobileNets as encoder to produce our masks.

Described in section 1.2, cycle consistency loss is introduced to further reduce random mappings. Seen from the Equation 4, the loss focuses on whole object, rather than the part. Therefore, we explicitly declare which part is important by specifying a mask. Firstly, we calculate the mask of the original image x , called M(x). The mask only highlights the hair as white and other part as black. Secondly absolute difference between the x and reconstructed one G(F(x)) is computed. Then since we only want to focus on the hair part, we apply an element-wise multiplication between mask M(x) and abs(G(F(x)) - x), followed by averaging. Hair masking will guide cycle consistency loss pay more attention to hair difference and ignore the other part of the face. Keeping the adversarial loss unchanged, we only modify the cycle consistent loss to the following:

$$\begin{aligned}\mathcal{L}_{\mathrm{cyc}}(\mathrm{G}, F) = {} & E_{x\sim p_{data}(x)}\left[abs\left(F\left(G(x)\right) - x\right) \cdot M(x)\right] \\ & + E_{y\sim p_{data}(y)}\left[abs\left(G\left(F(y)\right) - y\right) \cdot M(y)\right] \quad (5)\end{aligned}$$

3 Experiments

Our model is trained on Large-scale CelebFaces Attributes (CelebA) Dataset which contains more than 200k images with 40 attributes. Based on the blonde_hair and black_hair annotations, we divided the dataset into two, with black hair as domain X and blonde hair as domain Y. 75% is training set and the rest is testing set in each domain.

In the following, we compare our model with other unpaired image-to-image translation models, including the base model CycleGAN as well as DiscoGAN [6].

3.1 Qualitative measurement

During training time, we noticed that CycleGAN will focus more on reconstruction rather than translation in later epochs. From Figure 2, we can see the translated image in epoch 17 become less black in hair than epoch 16 in order to achieve better score in reconstruction loss. Also, DiscoGAN gives some uneven color blocks on face at 14th epoch, like the image shown in Figure 3.

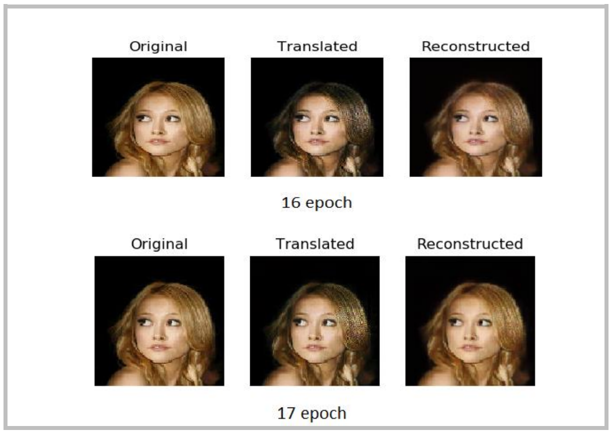

Figure 2: CycleGAN training

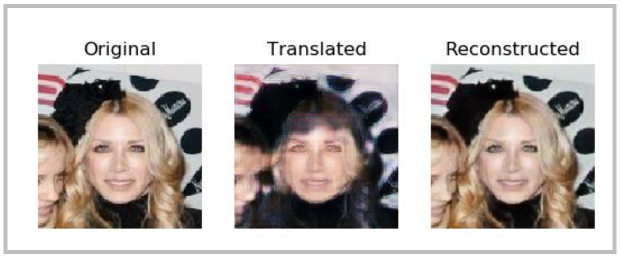

Figure 3: DiscoGAN 14 epoch

We select the 10th epoch from DiscoGAN, CycleGAN, and our model to make comparison since they all give reasonable results. As it can be seen from the Figure 4, our model changes the hair color attribute in more natural and uniform way. In CycleGAN transformed images, some highlight reflections on the blonde hair has been ignored, which causes an awkward boundary between transformed and untransformed part. For example, in the first row of Figure 4, CycleGAN translates part of the blonde hair to back but another part of the hair, which is the band, turns to pink. Comparatively, our model gives a better solution. It dyes the hair completely to black. Furthermore, in last row of Figure 4, DyeGAN can also distinguish the hair from the black hat on the head.

3.2 Quantitative measurement

Frechet Inception Distance (FID) score measures the similarity between original images and transformed images. FID converts the fake and real images into different gaussian distributions, fits the distributions to the hidden activations and then computes the Fréchet distance, also known as the Wasserstein-2 distance, between those Gaussians [7]. In general, a lower value in FID indicates a higher similarity between those two datasets. We randomly select 5000 pictures from our testing dataset, then score for DiscoGAN, CycleGAN, and DyeGAN are shown in Table 1:

Table 1: FID Score

	DiscoGAN	CycleGAN	DyeGAN
Blonde to Black	63.76765832196503	21.719315323853692	23.30066793165895
Black to Blonde	60.32030853542898	24.43403886101993	26.799733343163325

Seen from the above, DyeGAN gives a slightly higher score, which indicates the hair is dyed more completely since the translated images are evaluated as less similar than the original ones. DiscoGAN gives the highest score over the other two. After we compared the quality of three sets of images, this abnormal high score may be impacted by the noises, since this method is very sensitive to blur, swirl, black rectangular, or Salt and pepper noise. Images generated by DiscoGAN have a lower quality than the other two, which result in high fid scores [8].

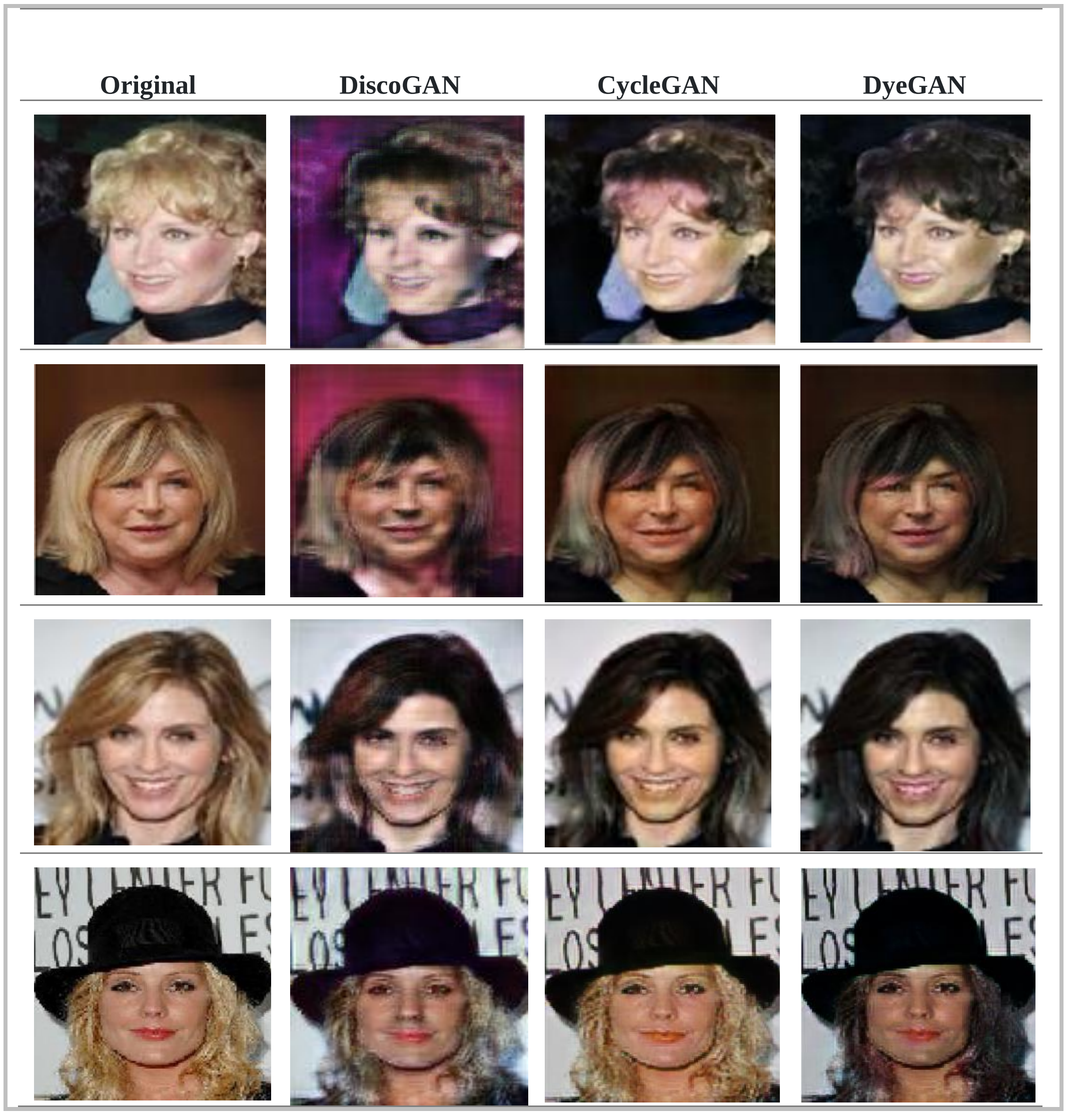

Figure 4: Comparison between three models in blonde hair to black hair translation

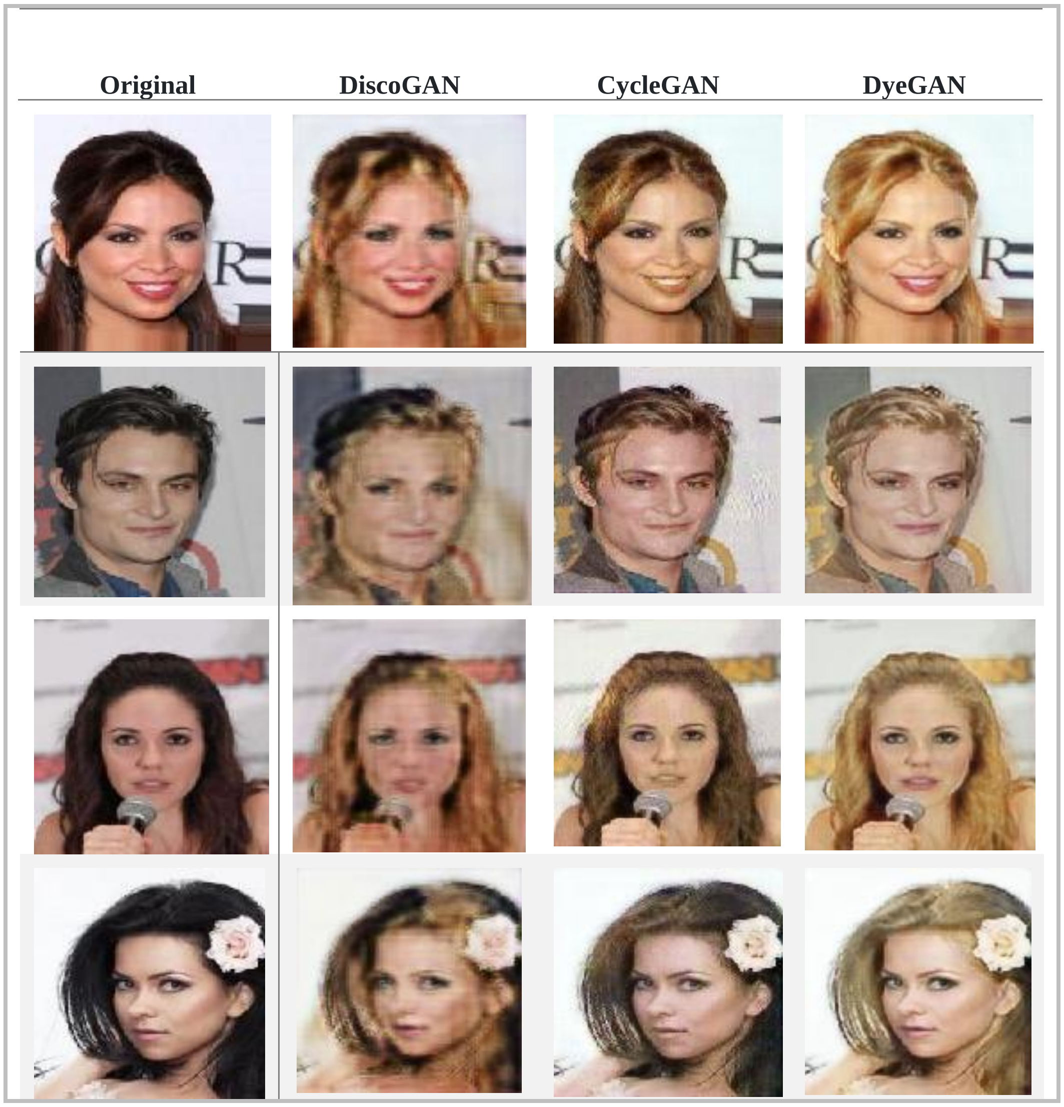

Figure 5: Comparison between three models in black hair to blonde hair translation

4 Conclusion

In this project, we build a new version of cycle-consistent adversarial networks, DyeGAN that can change the hair color under different circumstances. We guide our model only pay attention to the hair part by adding a segmentation mask to the reconstruction loss. As it can be seen from the above experiments, DyeGAN is able to change hair color in a more natural way. In later work, this segmentation idea can be applied to different parts of the face or even body and then easily change one of the attributes specified.

5 Contribution

Shiying Tu: In early research stage, Tu researched CycleGAN and AttGAN and also rebuilt runnable CycleGAN code in Keras. During implementation stage, Tu was responsible for training CycleGAN model and implementing DyeGAN based on Fully Convolutional Networks (FCN) model. For poster section, Tu assists to design the poster, and DyeGAN applications.

Lelei Zhang: In early research stage, Zhang researched DiscoGAN and StarGAN and also rebuilt runnable DiscoGAN code in Keras. During implementation stage, Zhang was responsible for training DiscoGAN and CycleGAN with sigmoid discriminators. Also, Zhang proposed to apply the idea of semantic segmentation to our project. For poster section, Zhang designs the poster.

Jui-Yang Yu: Yu provided the idea of working on CycleGAN and also did a research in the model in early stage.

The final report is primarily written by Lelei Zhang and Shiying Tu.

References

[1] Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycleconsistent adversarial networks. arXiv preprint arXiv:1703.10593, 2017

[2] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In CVPR, 2015.

[3] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. Retrieved August, 15:2018, 2018

[4] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015

[5] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. ArXiv: 1704.04861, 2017.

[6] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim. Learning to discover cross-domain relations with generative adversarial networks. In Proceedings of the 34th International Conference on Machine Learning (ICML), pages 1857–1865, 2017

[7] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In Proc. NIPS, pages 6626–6637, 2017.

[8] Ali Borji. Pros and Cons of GAN Evaluation Measures. In arXiv, pages 9-10,2018.