GENERATING IMAGES FROM SKETCHES USING GAN

Xiaoning Wan *Jun Hyek Jang* *Zhicheng He*
xw2501 *jj2883* *zh2318*

ABSTRACT

Generating images of face from sketches using a neural network is a challenging task. We implement several GAN models with bijective mapping to tackle this task. Specifically, DiscoGAN, CycleGAN and DualGAN are capable of learning cross domain with unpaired dataset, namely unsupervised learning, and we implement them to generate images from sketches. The result indicates that there is a clear improvement in generated images as we implement DualGAN and CycleGAN parameters on DiscoGAN.

***Index Terms*—** Face Synthesis, Unsupervised Learning, DiscoGAN, CycleGAN, DualGAN

1. INTRODUCTION

Generating a realistic image from a sketch has a wide range of applications. Within this topic, generating faces from sketches is one of the most challenging and useful topic. Previously, some GAN models, e.g. pix2pix [1], Context Encoders [2], have shown some promising results for this specific task.

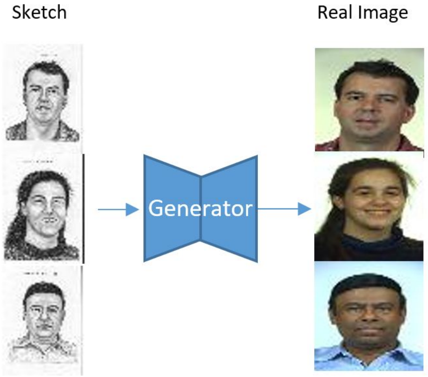

Figure.1 Generate Real Image from Sketch

The main drawback of these models is that they require strictly sketch-face paired dataset for training. To solve this problem, Xing Di et al [3]. introduced a method using conditional VAE [4] to generate sketch image from raw face image using unsupervised datasets. However, the generated sketch image is not satisfying enough, as there has been difficulties in optimizing loss functions in GAN models.

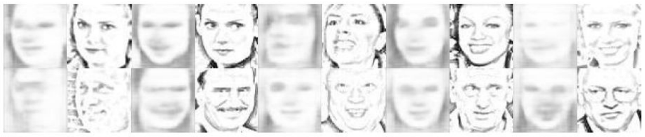

Figure.2 Samples of sketches generated by conditional VAE, odd columns show the generated sketches, even columns show the real sketches

To solve this problem, we studied Discovery GAN (DiscoGAN) [6], which uses shallow convolutional layers for generators and discriminators instead of a VAE. DiscoGAN can learn the cross domain of two domains, using unsupervised datasets.

In this paper, we analyzed DiscoGAN and implemented two other similar network, CycleGAN and DualGAN, on DiscoGAN to improve the result. We first test the DiscoGAN on 3D car dataset and Edges2Shoes dataset to verify its capability, and train it on Color FERET Database (sketch and face dataset). Then we improve DiscoGAN result by implementing loss functions of CycleGAN and DualGAN on DiscoGAN.

2. MODEL

In this experiment, we use several variant GAN models with a reconstruction loss to discover cross domain relation of two unsupervised datasets. Our goal is to train our models to learn bijective mapping between two domains, and we couple a pair of GAN with reconstruction loss to achieve this. In addition, unlike previous models of GANs, we do not use VAE for our generator and discriminator pair. Instead, we use shallow convolution layers for our generators and discriminators to take in images at the input instead of latent variables to aid discover the cross domain relation.

2.1. Formulation

Generator that takes in an image from domain A and generates an image in domain B is denoted as G_{AB}, and corresponding pair generator that takes in an image from domain B and generates an image in domain B is denoted as G_{BA}. The generated domain B image from domain A is denoted as $G_{AB}(x_A) = x_{AB}$.

In order to train a bijective mapping (one-to-one correspondence) in generators, we constraint the pair of generators to be an inverse of each other. In addition, we calculate the reconstruction loss between the input image x_A and output image of $G_{BA}(G_{AB}(x_A))$ for each corresponding pair of GANs. The model structure for GAN with a reconstruction loss is shown in Figure3. By optimizing the reconstruction loss, each generator can learn the mapping from its input domain to output domain and discover the relations between them. The equation of reconstruction loss is shown in the below equation.

$$L_{constA} = d(G_{BA}(G_{AB}(x_A)),\ x_A) \qquad (1)$$

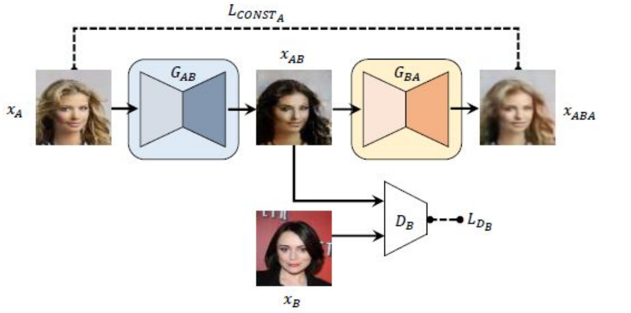

Figure.3 Model of GAN with a Reconstruction Loss

The discriminator that discriminates images in domain A is denoted as D_A. Similarly, the discriminator that discriminates images in domain B is denoted as D_B.

The total loss for each generator G_{AB} is just a addition of the reconstruction loss and GAN loss from a discriminator D_B with input of x_A, and the equation is shown below.

$$L_{GAN_B} = -E_{x_A \sim P_A}[log D_B(G_{AB}(x_A))] \qquad (2)$$

$$L_{G_{AB}} = L_{GAN_B} + L_{CONSTA} \qquad (3)$$

The discriminators use the standard GAN discriminator loss first described by Goodfellow et al [5].

$$L_{D_A} = -E_{x_A \sim P_A}[log D_A(x_A)] \qquad (4)$$

2.2. DiscoGAN

Discovery GAN (DiscoGAN) is a model that was first proposed by Kim et al [6] in 2017. GAN with a reconstruction loss can only learn unijective mapping where the generator can only generate images in domain B from domain A. DiscoGAN learns bijective mapping by coupling two GANs with reconstruction loss together to train unsupervised datasets of two domains. The model structure is shown in figure 4.

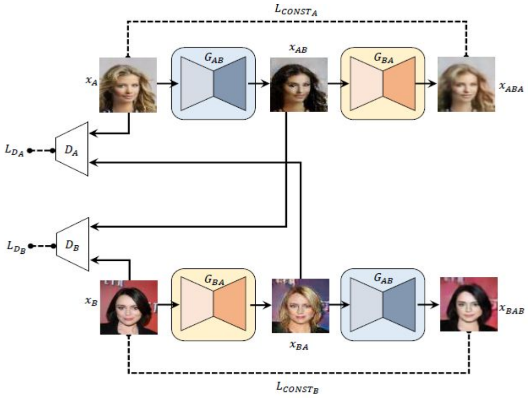

Figure.4 Model of DiscoGAN

The cross domain training is kept under constraint via the the Reconstruction Loss and the coupled discriminator loss. This model has similar generator and discriminator losses as the GAN with a reconstruction loss.

$$\begin{aligned} L_{D_A} = & -E_{x_A \sim P_A}[log D_A(x_A)] \\ & -E_{x_B \sim P_B}[log(1 - D_B(G_{AB}(x_A)))] \end{aligned} \qquad (5)$$

$$L_G = L_{G_{AB}} + L_{G_{BA}} \qquad (6)$$

2.4. CycleGAN

CycleGAN is a variation of DiscoGAN with the same model structure, but different generator loss function. In the generator loss, CycleGAN has an additional cyclic consistency term. The generator loss function is described below.

$$L_{G_{AB}} = L_{GAN_B} + L_{CONSTA} + L_{cyclic} \qquad (7)$$

$$L_{cyclic} = d(G_{AB}(x_B),\ x_B) \qquad (8)$$

The cyclic loss is just a distance between $G_{AB}(x_B)$ and x_B where it often referred to as the identity loss. Theoretically, $G_{AB}(x_B)$ should be unchanged if the input image is from domain B. The cyclic loss term adds further constraints on the generator to be cyclically consistent between G_{AB} and G_{BA}, and learn the bijective mapping more efficiently.

2.5. DualGAN

DualGAN has the same structure and loss function as CycleGAN, but with an additional constant hyperparameter λ that is multiplied to the generator and discriminator loss function.

$$L_{G_{AB}} = L_{GAN_B} + L_{CONSTA} + \lambda_{cyclic} L_{cyclic} \tag{9}$$

$$\begin{aligned} L_{D_A} = & -\lambda_D E_{x_A \sim P_A}[log D_A(x_A)] \\ & -\lambda_D E_{x_B \sim P_B}[log(1 - D_B(G_{AB}(x_A)))] \end{aligned} \tag{10}$$

The hyperparameter helps training by putting more emphasis on the cyclic loss in the generator function, which makes the model more cyclically consistent.

In addition, DualGAN uses 8 layer Unet as generator but the same discriminator as CycleGAN.

3. EXPERIMENT

We first start with testing DiscoGAN on simple datasets of Car to Car translation and Sketch to Shoes. Then, we train DiscoGAN, CycleGAN and DualGAN on Sketch to Face dataset.

In each experiments, we set the input size to be $64 \times 64 \times 3$, which is same to the original setting of DiscoGAN. We set the epoch size to be 5000 and batch size to be 64. For our optimizer, we choose Adam optimizer with $\beta_1 = 0.5$ $\beta_2 = 0.999$. The default learning rate = 0.0002 and we decrease the learning rate every 600 iterations. All the datasets of domain A and B are unsupervised pairs. We train/test the model using Google Cloud Platform using GPU NVIDIA Tesla K80.

3.1. Car2Car - DiscoGAN

In this experiment we use our model to translate 3D car model into exact opposite angle. We trained and tested our model on 3D car model data from (Fidler et al., 2012), which contains 199 3D car model in 12 directions, each direction varying 15° from the adjoining images.

3.2. Sketch2Shoes - DiscoGAN

In this experiment we use our model to translate sketches of shoes into images of shoes. The dataset we use in this experiment are from UT-Zap50K. It each contains 49444 color images of shoes photographs and shoes sketches. In order to evaluate our model, we set sketch to be domain A and photograph image to be domain B.

3.3. Sketch2Face - DiscoGAN

In this task, we trained and tested our model on Color FERET Database, which contains 1194 color images of 1194 people and its corresponding sketches. Unlike 3D car model and shoes, human faces have much more details and less number of datasets. Therefore, we increase the epoch to 10 000 and adjust learning rate every 1000 iterations.

3.4. Sketch2Face - DualGAN

Using the same FERET Dataset, we implement DualGAN's generator and loss function on DiscoGAN. We increase the generator depth to 8 layers, and set the hyperparameter, λ, to 10 in the loss function.

3.5. Sketch2Face - CycleGAN

Using the same FERET Dataset, we have added CycleGAN's cyclic loss term as described in eq.8 into the DiscoGAN's generator loss function.

4. RESULT AND DISCUSSION

4.1. Car2Car - DiscoGAN

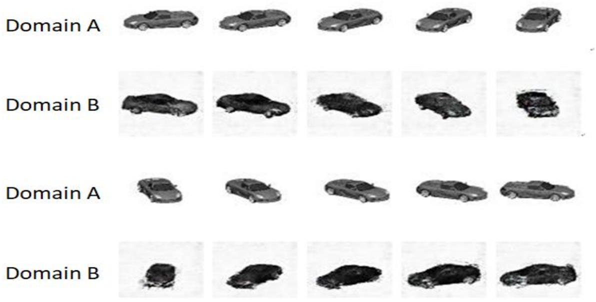

Figure.5 Car to Car domain experiment

DiscoGAN shows a considerably good performance in the car to car experiment. The Generated 3D car model are in the opposite angle of the origin model, which indicates that the DiscoGAN model is learning bijective mapping.

4.2. Sketch2Shoes - DiscoGAN

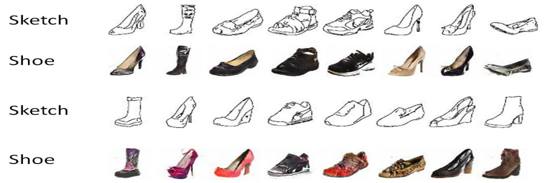

Figure.6 Sketch to Shoes domain experiment. Image of shoes is generated using a sketch of shoes

The DiscoGAN successfully generated color images of shoes from grayscale images of sketch. This indicates that DiscoGAN can learn the bijective mapping of two unpaired domains that have different complexity (color and grayscale).

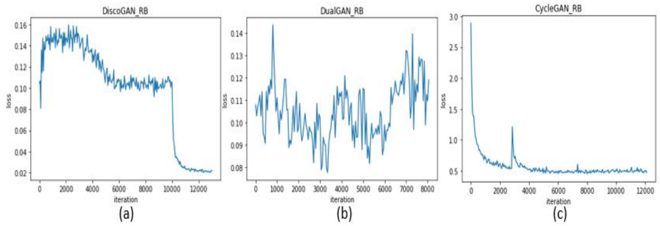

Figure.7 (a) Reconstruction Loss of DiscoGAN (b) Reconstruction Loss of DualGAN (c) Reconstruction Loss of CycleGAN

4.3. Sketch2Face - DiscoGAN

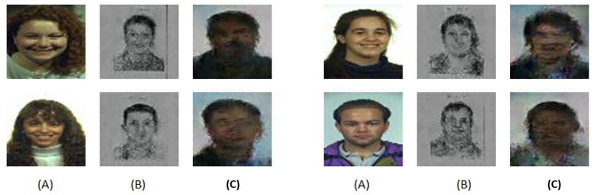

Figure.8 (A) Input face images x_A (B) Sketches generated by generator $G_{AB}(x_A)$ (C) Reconstructed face image generate by $G_{AB}(G_{BA}(x_A))$

DiscoGAN failed to produce a satisfying result in this task as reconstructed face is deformed. DiscoGAN has hard time optimizing the reconstruction loss as seen in Figure7. To improve this, we implement Cyclic constraints from CycleGAN and DualGAN in the following experiments.

4.4. Sketch2Face - DualGAN

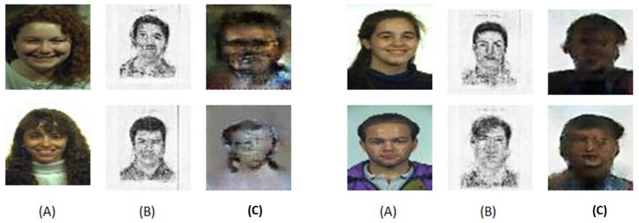

Figure.9 (A) Input face images x_A (B) Sketches generated by generator $G_{AB}(x_A)$ (C) Reconstructed face image generate by $G_{AB}(G_{BA}(x_A))$

With deeper generator structure and emphasized cyclic loss implementations, the result is worse than the DiscoGAN's result. It seems like we put too much constraints on the cyclic consistency and the reconstruction loss starts to overshoot as seen in Figure.7. To reduce overshoot, we modified our loss function based on CycleGAN.

4.5. Sketch2Face - CycleGAN

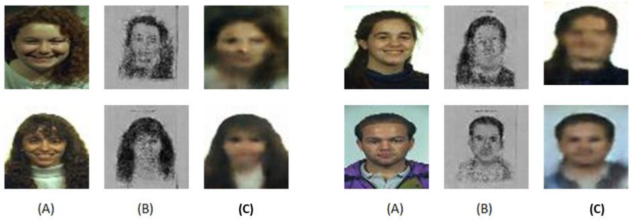

Figure.10 (A) Input face images x_A (B) Sketches generated by generator $G_{AB}(x_A)$ (C) Reconstructed face image generate by $G_{AB}(G_{BA}(x_A))$

Adding a cyclic term in DiscoGAN's generator loss improved both the regenerated faces and the reconstruction loss. This model produces regenerated silhouette of faces without any deformation at the boundary as in Figure10. The reconstruction loss is optimized monotonically like an ideal convolution networks. The limitation of regenerating details on the face maybe due to small input size of 64 X 64 X 3 as a lot of details are lost during downsampling process.

5. CONCLUSION

In this work, we explore DiscoGAN, DualGAN and CycleGAN that learns bijective mapping on unsupervised datasets to generate faces from sketches. In the case of DualGAN, we discover that putting too much consistency constraints leads to overshooting of reconstruction loss, and in the case of DiscoGAN, putting too little constraints leads to slow convergence of reconstruction loss. There is a perfect 'Goldilock Zone' of consistency constraints that produces the optimal result.

In the future, we can further improve the face synthesis from a sketch by modifying the weight initialization and deeper generator and discriminator structure.

6. REFERENCES

[1] Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017)

[2] Pathak, D., Kr¨ahenb¨uhl, P., Donahue, J., Darrell, T., Efros, A.: Context encoders: Feature learning by inpainting. In: CVPR (2016)

[3] Xing Di,Vishal M. Patel, A.: Face Synthesis from Visual Attributes via Sketch using Conditional VAEs and GANs. arXiv preprint arXiv:1801.00077

[4] Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. In: Advances in Neural Information Processing Systems, pp. 3483–3491 (2015)

[5] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In Advances in Neural Information Processing Systems (NIPS), 2014.

[6] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, Jiwon Kim ; Proceedings of the 34th International Conference on Machine Learning, PMLR 70:1857-1865, 2017.