Improving Learning time in Unsupervised Image-to-Image Translation

Tae-Hong Min, Do-Yun Kim, Young-June Choi

Department of Software

Ajou University

Suwon, The Reupublic of Korea

carpediem804 @ajou.ac.kr

***Abstract*— Unsupervised image-to-image translation can map local textures between two domains, but typically fails when the domain requires big shape changes. It is difficult to learn how to make such big change using the basic convolution layer, and furthermore it takes much time to learn. For faster learning and high-quality image generation, we propose to use Cycle GAN that is combined with Resnet in a network that is connected with the residual block for upsampling to make big shape change and construct faster image-to-image translation.**

Keywords—GAN, CNN, deep learning, DiscoGAN

I. INTRODUCTION

Unsupervised image-to-image translation is the process of learning any mapping between labels or unpaired image domains. This process is usually implemented using generic adversarial networks (GAN) using hostile learning between the discriminator network and the generator network, and cyclic loss to overcome the lack of supervised pairing.

DiscoGAN [2] and CycleGAN [3] enable delicate representation between image domains, such as transforming pictures. Especially, CycleGAN can be used for conversion to high quality image, but the conversion is known to be difficult. A network with fully connected generators such as DiscoGAN may show a larger type of change if sufficient network capacity is given. Otherwise, the image quality is deteriorated.

In this study, we applied the concept of residual block of resnet to the generator network part of DiscoGAN to improve the learning speed and image quality of the generated image while improving the image quality of existing DiscoGAN.

II. Related Work

A. Generative Adversarial Networks

As shown in Figure 1, Generative Adversarial Networks (GAN) [1] has employed an adversarial loss function of generators and discriminators to distinguish real images from generated images, thus generating many results in image editing and image synthesis. We also show that GAN can learn texture mapping between complex domains through Pix2Pix. As shown in Figure 2, however, this technique requires a large number of paired sample data. In addition to simply adding a loss function via CNN, adding a Gan loss produces a photo-realistic image. In other words, if DCGAN learns to extract the data distribution from noise distribution, Pix2Pix learns a mapping function between two image domains. Through the discriminator, it is possible to check whether the generated one is real or not

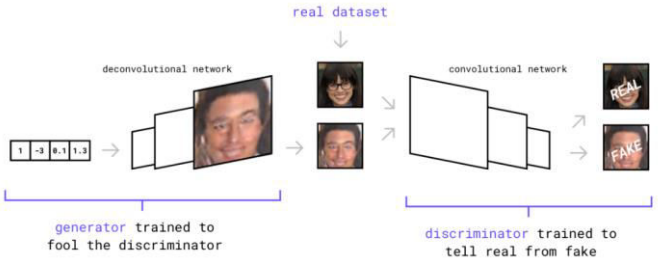


Figure 1 Gan architecture

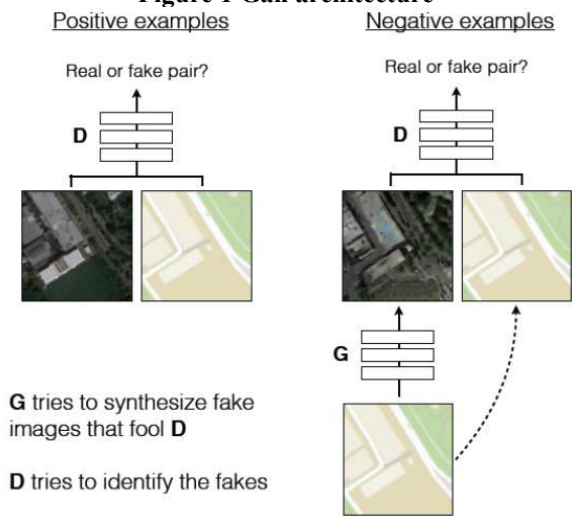


Figure 2 Pix2Pix architecture

B. Unsupervised Image Translation GANs

Unsupervised Image Translation GANs [3] extended learning techniques such as Pix2Pix[4] to enable learning in unsupervised pairs. When the image domains X and Y are given, learning of cyclic mapping in X-> Y-> X and Y-> X-> Y is performed to prevent the mode collapse in the unsupervised case, as depicted in Figure 3.

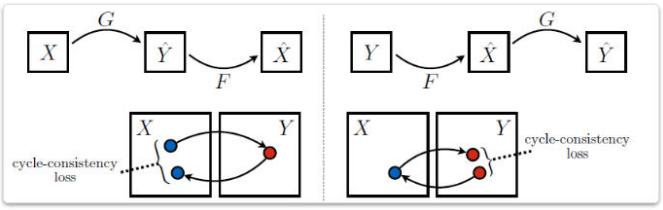


Figure 3 Cyclic mapping

Such an architecture is adopted in CycleGAN and DiscoGAN. CycleGAN uses resnet or U-net in the generator network to make it easier to learn and generate high-quality images with a fast learning time. However, because of the influence of the previous layer, the format conversion tends to be difficult. Since DiscoGAN is of an encoder-decoder format, there is no influence between layers in the Generator Network. Therefore, it is possible to convert the format easily. However, it is difficult to convert the image of high quality.

III. Our Approach

In unsupervised image-to-image translation, when the shape of data images is very different, the encoder-decoder format is used for learning. However, there is a problem that it takes some time to learn because there is no influence between layers. The main idea of Resnet's idea is, as shown in Figure 4, to use a residual block that creates a skip connection so that the gradient can flow well. This is similar to the Long Term Short Memory (LSTM), which introduces a forget gate and so on to better flow the gradient of the previous step. In this study, by using these residual blocks in the generator upsampling, it is possible to flow the big feature well, but it does not have big shape transformation, and it is easy to learn because the image quality is improved, the network is structured with high speed. and a generator with residual blocks is constructed.

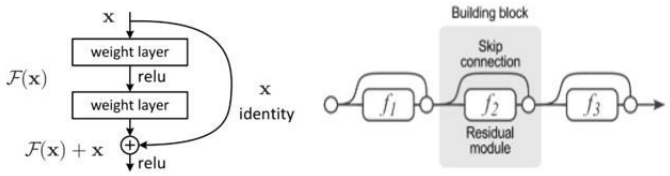


Figure 4 Resnet skip connection architecture

The architecture of this study is based on DiscoGAN and CycleGAN. The existing DiscoGAN has the basic encoder-decoder structure as shown in Figure 5, and it is easy to change the format. However, since it has the basic decoder structure of the network, it is difficult to learn upsampling through the narrow bottleneck layer and output image size. And it has a low resolution limit due to the low capacity of the network. Therefore, the generator in this study includes residual bocks in the upsampling layer, thus making it easier to learn and be capable of producing images of high quality.

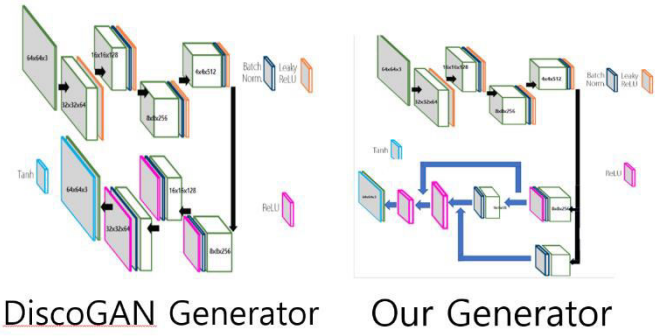


Figure 5. Comparison of generator architectures

In addition, to connect the residual block in the encoder part, that is, downsample in the encoder-decoder, it takes a long time and a mode collision occurs. The reason is that if one uses the Residual block in the encoding part, it is more difficult to learn when converting to other image when upsampling is done by learning the whole existing input image. However, in this paper, when the residual block is used in the decoder part, that is, in the upsampling part, the residual image is obtained by extracting the features of the existing input image and using the residual block

IV. EXPERIMENT RESULTS

In this study, we use facecrubs, which consists of 100,000 face images of 530 people, as image-to-image data set. A

total of 500 epochs are used and GPU is 1080Ti. The learning time of DiscoGAN is 61 hours and 41 minutes to reach 500 epochs. In the network of this study, it took 41 hours and 54 minutes. It is faster by 33% than existing DiscoGAN.

In order to determine the morphological transformation, we checked the probability of the transformed data through CNN. As a result, DiscoGAN showed 8.77%, while our method showed 8.92%. Therefore, it can be confirmed that there is no difference in morphological transformation. .

Table 1 Total learning time and Morphological transformation error

	Ours	DiscoGAN
500 Epoch Time	**41:54:60**	**61:41:44**
Classification Error	**8.92%**	**8.77%**

Figures 6 and 7 show images generated by DiscoGAN and the proposed method. There is no big difference between both methods in many cases, but our method sometimes shows better image quality than DiscoGAN.

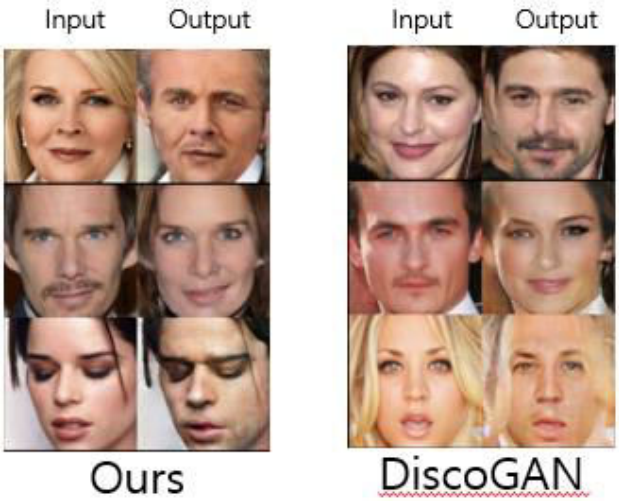
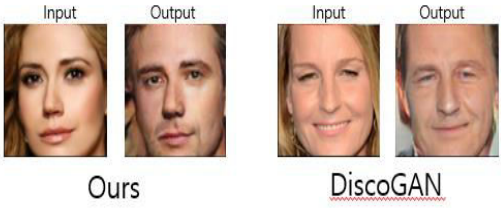


Figure 6 DiscoGAN and Our result images



	OURS	DISCOGAN
PSNR	18.014740	17.3908
SSIM	0.722216878	0.7291176
MSE	1027.084779	1185.766061
RMSE	32.048	34.434

Figure 7. Resolution Result

V. CONCLUSION

In this study, we used a residual block in the upsampling structure in the generator to maintain the shape transformation and construct a network that takes much less time to generate high quality images. One of the problems in existing unsupervised image-to-image translation, which takes a long learning time, is solved by using a residual block, and the image can be converted at a higher image quality than the existing research. In future work, we plan to construct a more efficient network by combining new ideas from generator or loss function as well as generator structure.

ACKNOWLEDGMENT

"This research was supported by the MIST(Ministry of Science and ICT), Korea, under the National Program for Excellence in SW supervised by the IITP(Institute for Information & communications Technology Promotion)" (2015-0-00908)

REFERENCES

[1] Radford, A., Metz, L., Chintala, S.: Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. ArXiv e-prints (Nov 2015)

[2] Kim, T., Cha, M., Kim, H., Lee, J.K., Kim, J.: Learning to discover cross-domain relations with generative adversarial networks. In: International Conference on Machine Learning (2017)

[3] Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networkss. In: International Conference on Computer

[4] Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Computer Vision and Pattern Recognition (2017)

[5] Liao, J., Yao, Y., Yuan, L., Hua, G., Kang, S.B.: Visual attribute transfer through deep image analogy. ACM Trans. Graph.

[6] Aaron Gokaslan, Vivek Ramanujan, Daniel Ritchie, Kwang In Kim, and James Tompkin. Improving shape deformation in unsupervised image-to-image translation. arXiv preprint arXiv:1808.04325, 2018.