Make up for Anime Character with CycleGAN, BeautyGAN and U-GAT-IT

Yijia Chen¹,a,† and Rui Gong²,b,†
¹Department of Information Management, Chung, Yuan Christian University, Taoyuan, Taiwan, 20214, China
²Department of Computer Science and Technology, Tsinghua University, Beijing,100085 China
a10344270@cycu.org.tw, bgongr17@mails.tsinghua.edu.cn
†These authors contributed equally.

Abstract. Image-to-Image translation is a class of problems which aim to translate images from domain A into domain B. Some methods have been proposed to solve these problems, each of which being skilled in some specific tasks, e.g., CycleGAN for translating images into Monet style, U-GAT-IT for translating selfies into anime characters. This paper discusses how to finish an A→B task through combining an A→C approach and a C→B approach, and carries out experiments on generating anime characters with makeup from selfies combining CycleGAN, BeautyGAN and U-GAT-IT. The result shows that sequence is a key factor in our approach, and the reason why different sequences produce different results is analyzed.

1. Introduction
Image to image translation is a method to convert the image feature into another image feature. In other word, we need to find a function that allows the A domain image to be mapped into the B domain, which can be used to solve many problems, such as Style transfer, attribute transfer, improvement of image resolution.

There is a common idea in the early image style transfer program, which is to analyze a certain style of image, build a model for that style, and then change the image to be transferred. Although the effect is good, it also brings a drawback that cannot be ignore: a program can basically only be used in a certain scene or a certain style. Thus, practical applications based on traditional style transfer research are very limited.

In order to solve this limitation. Leon A. Gatys et al. use CNN (Convolutional Neural Networks) to finish the task [1], that is, combining the style image’s style with the content image’s content to generate a new image. But this method still has some drawbacks. Because the original input is actually a bunch of Gaussian white noise, through constant BP adjustment, an image will eventually be formed. If we change an image for rendering, we still have to start from scratch. It can be said that the final image can be obtained through the continuous comparison and correction of the input value. Therefore, online, real-time drawing cannot be achieved.

Li-feifei et al. introduce GAN (Generative adversarial networks) into this field [2]. It obviously improved the efficiency of the image-to-image translation. They turn the task of style transfer into a single generation process by using Image Transform Net as generator and VGG-16 as discriminator.
Among them, generator is used to fake pictures. And discriminator is used to extract style and content information as training of generator.

Mirza, Mehdi, and Simon Osindero raised C_GAN (Conditional Generative Adversarial Networks) in 2014 [3]. In C-GAN, both the generator and the discriminator add supernumerary information y as a condition, and y can be any type of supernumerary information, like category information or data from other modalities.

The Pix2Pix model, proposed by Isola, Phillip, et al [4], is based on C-GAN and adopts an end-to-end architecture, where jump connections are introduced in the generator architecture to preserve the image latent layer structure. Original image x is the input. And the output is translated target image G(x). The domain image and the true or false target domain image are respectively combined as the discriminator’s output, and the discriminator outputs the classification result and confronts the generator. The problem with Pix2Pix is that one-to-one mapping between the original domain and target domain is used when training the model, in which leading to poor model diversity. In order to resolve the drawback, Jun-Yan Zhu et al. proposed BicycleGAN [5]. BicycleGAN adopted Introduce latent layer coding to restrain the output’s bijective consistency and latent layer coding to improve the diversity of models

Both of above method are supervised image-to-image translation. One problem with supervised image translation models is that there are not so many paired data sets to use. Therefore, more and more unsupervised image-to-image translation models have been proposed recently, one of which is CycleGAN [6]. It’s idea is to design a cycle consistency, and use it to supplant the reconstruction loss.

Kim, Junho, et al. modified the generator and discriminator of CycleGAN and proposed U-GAT-IT [7] which can convert images of real people to anime characters.

Li, Tingting, et al. synthesized the global domain-level loss and the local instance-level loss, and used two GAN networks to created BeautyGAN [8] which can make up for the portrait.

Recent studies have proposed that the third network, Siamese, can be added to replace the loop consistency loss, thereby reducing the model’s complexity and training costs. To learn the high-level semantic features of the image Siamese network is used which can obviously improve the similarity between translated image and original image. According this concept, TraVeLGAN [9] was proposed by Matthew Amodio and Smita Krishnaswamy. Another type of image translation model is to encode the content and attributes of the global image, and realize image translation by exchanging the attribute codes. Among them, the representative models are DRIT proposed by Hsin-Ying Lee et al. [10], and MUNIT proposed by Xun Huang et al. [11].

StarGAN is proposed by Y. Choi, M. Choi et al. [12] to solve training across multiple domains and multiple data sets. In StarGAN, traditional fixed translation is not used, but domain information and pictures are input together for training, and a mask vector is added to the domain label to facilitate different training sets to joint training.

Dongwook Lee et al. proposed a new missing image data interpolation framework called CollaGAN[13]. CollaGAN researches multi-domain image-to-image translation tasks based on the now mature single-image-to-image generation, so that a single generator and discriminator network can use the remaining clean data set to successfully estimate the missing data.

Currently, more and more models have been introduced to guide image translation by introducing attention mechanisms, such as Selection GAN proposed by Hao Tang, Dan Xu [14], CSA proposed by Hongyu Liu et al. [15] and so on.

Make-up always is an important thing to people who love good-looking, and when the invention of the camera, people have come up with a lot of methods to modify the images and make-up it, at the very beginning, it may be a handcraft work, and after the popularization of the digital camera, many application has been made to make up the portrait, the most simple way is to add a filter, and when we have had the technology of machine learning, we can use some method like BeautyGAN to do the make-up work, and because there are some technologies like U-GAT-IT make us can created anime character, so our team have an idea to combine the BeautyGAN and U-GAT-IT to create a method to make up for the anime characters. Just like make up for real people, make up for anime characters also can bring
some amusement function, and I believe it can attract people who are interesting in the animation and cartoon.

To achieve our goal, we combined three kinds of GANs: CycleGAN, BeautyGAN and U-GAT-IT. We implemented the pre-trained model on selfie dataset to generate anime characters with make-up.

Method

As shown in Figure 1, our work is based on CycleGAN, BeautyGAN and U-GAT-IT and followed by this process:

1. Using CycleGAN to transfer real image to some artistic style (we find Monet style is best),
2. Using BeautyGAN to make it up.
3. Using U-GAT-IT to create a fictional image.

1.1. CycleGAN.
CycleGAN was raised by JunYan Zhu et al. in 2017. CycleGAN uses two mirror-symmetric GANs, using it to build a loop network. Both GANs share their generators, and each has a discriminator, which means, there are two discriminators and two generators. A one-way GAN has two losses, so CycleGAN has four losses. The advantage of CycleGAN is it can train two image sets without pairing.

1.2. BeautyGAN.
In 2018, Ting ting Li et al. proposed a method to make up for the given portrait, it was called BeautyGAN. BeautyGAN learn the mapping between two domains at the same time. Which means, For the source image ∈ A domain and reference image ∈ B domain, BeautyGAN can generate post-make-up image ∈ B domain and anti-make-up image ∈ A domain.

BeautyGAN has a generator G and two discriminators DA and DB. Two input images from domain A and domain B will be sent to the generator G, and then G will translate those images into output: post-make-up image and anti-make-up image. Then, they will be fed into the generator G and it will bring a reconstruction results: rec-source image ∈ A and rec-reference image ∈ B.

1.3. U-GAT-IT.
In 2019, J Kim et al. came up with two methods to improve the unsupervised image-to-image translation and improve the style transfer on the large-scale. For example, to change the portrait to the anime style.

The first method is introducing Explainable AI such as Class Attention Maps (CAM) into the GAN and utilized L1 norm and LSGAN to design the CAM loss. The second method is to invent AdaLIN, and set the weight to let the network know they should use Instance Norm or Layer Norm.

U-GAT-IT is based on CycleGAN, which added CAM and AdaLIN. For Discriminator, it just added CAM.

There are two generators: G {s→t, t→s} and two discriminators: D {s, t} In U-GAT-IT, beside GAP, Global Maximum Pooling (GMP) will be used also. Thus, the neural network will notice the small area. The outcome from GAP and GMP will concatenate to each Encoder Feature Map to multiply.
2. Experiment results
We apply three kinds of GANs: CycleGAN, BeautyGAN and U-GAT-IT with pretrained model on selfie dataset in different order to generate anime characters with makeup of similar style with a given image.

First, we tried using U-GAT-IT to generate anime characters from selfies and then use BeautyGAN to make up the anime characters. The results are anime-styled, but BeautyGAN failed to make them up due to the difference between the source domain BeautyGAN is trained on and now applied on. Result is shown in Figure 2. [0,0] source image, [0, 1~9] represent different makeup styles, [1,0] represents result of BeautyGAN, [1, 1~9] represent final result.

Then the order of GANs was changed to BeautyGAN, CycleGAN and U-GAT-IT, making selfies up first and then transferring them into Monet Style and finally generating anime characters. The results showed difference in lips rouge corresponding to different kind of makeup, but the areas of lip rouge do not fit mouths of anime characters well. Result is shown in Figure 3. [0, 1~9] different makeup styles, [1, 0] source image, [1, 1~9] result of BeautyGAN, [2, 1~9] result of CycleGAN, [3, 1~9] final result.

The order of GANs of CycleGAN, BeautyGAN and U-GAT-IT was also experimented, which gives a better result. Result is shown in Figure 4. [0, 0] source image, [0, 1~9] different makeup styles, [1,0]
result of CycleGAN, [1, 1–9] result of BeautyGAN, [2, 1–9] final result.

3. Analysis
Although the total loss increases though each ‘layer’, the U-GAT-IT part may contribute the greatest to the total loss. The U-GAT-IT part and CycleGAN part transfer real world style to non-real styles, while the BeautyGAN transfer real world style to real world style. As result, the loss of concatenating BeautyGAN and U-GAT-IT is less than concatenating CycleGAN and U-GAT-IT. And the result is more natural.

4. Conclusion
In order finish the task of translating selfies into anime characters with makeup, we experimented different kinds and combination of GANs. Finally, it is found that combining CycleGAN, BeautyGAN and U-GAT-IT is an effective approach. And the sequence of these GANs is a key factor to improving performance. We proposed that reducing difference between domains at each step of translation is an effective method to achieve a better result.

References
[1] Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In: IEEE conference on computer vision and pattern recognition. Las Vegas. pp. 2414-2423.
[2] Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. Amsterdam. pp. 694-711.
[3] Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784. https://arxiv.org/abs/1411.1784
[4] Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In: IEEE conference on computer vision and pattern recognition. Honolulu. pp. 1125-1134.
[5] Zhu, J. Y., Zhang, R., Pathak, D., Darrell, T., Efros, A. A., Wang, O., & Shechtman, E. (2017). Toward multimodal image-to-image translation. arXiv preprint arXiv:1711.11586. https://arxiv.org/abs/1711.11586
[6] Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE international conference on computer vision. Venice. pp. 2223-2232.
[7] Kim, J., Kim, M., Kang, H., & Lee, K. (2019). U-GAT-IT: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. arXiv preprint arXiv:1907.10830. https://arxiv.org/abs/1907.10830
[8] Li, T., Qian, R., Dong, C., Liu, S., Yan, Q., Zhu, W., & Lin, L. (2018). Beautygan: Instance-level facial makeup transfer with deep generative adversarial network. In: 26th ACM international conference on Multimedia. Seoul. pp. 645-653.
[9] Amodio, M., & Krishnaswamy, S. (2019). Travelgan: Image-to-image translation by transformation vector learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach. pp. 8983-8992.
[10] Lee, H. Y., Tseng, H. Y., Huang, J. B., Singh, M., & Yang, M. H. (2018). Diverse image-to-image translation via disentangled representations. In: European conference on computer vision (ECCV). Munich. pp. 35-51.
[11] Huang, X., Liu, M. Y., Belongie, S., & Kautz, J. (2018). Multimodal unsupervised image-to-image translation. In: European conference on computer vision (ECCV). Munich. pp. 172-189.
[12] Choi, Y., Choi, M., Kim, M., Ha, J. W., Kim, S., & Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: IEEE conference on computer vision and pattern recognition. Salt Lake City. pp. 8789-8797.
[13] Lee, D., Kim, J., Moon, W. J., & Ye, J. C. (2019). CollaGAN: Collaborative GAN for missing image data imputation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach. pp. 2487-2496

[14] Tang, H., Xu, D., Sebe, N., Wang, Y., Corso, J. J., & Yan, Y. (2019). Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach. pp. 2417-2426.

[15] Liu, H., Jiang, B., Xiao, Y., & Yang, C. (2019). Coherent semantic attention for image inpainting. In: IEEE/CVF International Conference on Computer Vision. Seoul. pp. 4170-4179.