計算工学講演会論文集　Vol.28 (2023年5月)　　計算工学会

Generating the human body structure of the corresponding anime character based on DiscoGAN

Liu Sihan [1)], Ryuji Shioya [2)] and Yasushi Nakabayashi[2)]

1) Graduate School of Information Sciences and Arts, Toyo University（350-0815 2100 Kujirai Kawagoe Saitama,

E-mail: ishlauin@gmail.com)

2) Faculty of Information Sciences and Arts, Toyo University

With the popularity of social networking services and smartphones, photo processing applications have become widely used, and there is a growing interest in more advanced photo processing techniques. In the field of image generation, methods based on Generative Adversarial Networks (GANs) have shown particularly superb results. We discussed Generating the human body structure of the corresponding anime character based on GANs. This research is expected to be applied in the field of drawing that can be used as a tool to significantly reduce learning period or provide a new study method for drawing beginners.

Key Words: *GANs, generating, body structure, anime character*

1. INTRODUCTION

Over the past few years, many researchers have shown an interest in assisting painting based on deep learning. Even though there is much research have adopted the approach that fully automatic [1][2] and semi-automatic [3][4] supplementary methods, either method has little effect on efficiency or cost. The problem seems to lie in the fact that they always focus on colorization or directly generate a character and image. However, there has been little study done concerning how to help learners to draw a good sketch, and beginners always spend too much time and energy on learning how to draw a correct human body structure of character. In order to help beginners to gain more experience in less time, it is important to find a fully automatic way to human body structure extraction of anime characters. The main objective of this thesis is to build a model to automatically recognize the pose of the anime characters in the illustration, through this pose [5] to generate the corresponding human body structure and make sure they could get a correct reference rapidly.

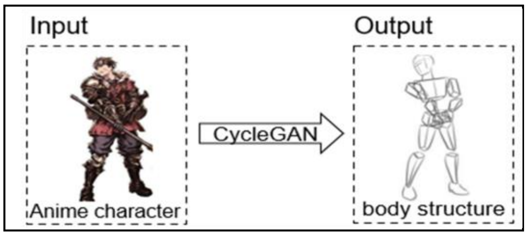

Fig. 1: The flow with using CycleGan

2. PURPOSE

Although the ultimate purpose of this research is to help drawing beginners get a faster grasp on how to draw a correct anime style character’s body structure, the current purpose is to find a way to quickly and accurately extract the specified style of body structure from anime characters, therefore, the flow assumed as Fig. 1. Meanwhile, most pioneer researchers generally believe that it is more valuable to keep the content in the original image unchanged and make changes to its style or translate it. And it is meaningless to research if the output results in random content. This view leads to less prior research on how to use deep learning models to modify the content of images, so how to control the output results within a certain range is one of the difficulties in this research.

3. MODEL

In this section, we propose a way how to convert an anime character image to a simple body structure image. The current dataset for this study was collected on the web, but unfortunately, we could not find a website that specifically stored images about anime style human body structures, which led to a difficult collection of paired datasets. While pix2pix [8] is very powerful in image generation and transformation, and the result is always satisfactory, it is not suitable for this study due to its high requirements on the training set (needs pairs of images). Unsupervised model CycleGan [9] does not use paired image datasets as shown in Fig.2, it for the most part solves the problem of difficult collection of data sets.

Although we will use an unpaired dataset in this research, the original CycleGan neural network will only change the style of the image, not the content of the image, so the loss function needs to be reset or changed to allow the model to modify the content of the image. By changing the loss function and adjust the model to achieve the desired results.

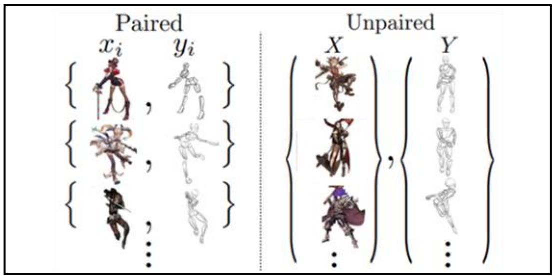

Fig. 2: Paired training data (left) consists of training examples $\{x_i,\ y_i\}$ N_i=1, where the y_i that corresponds to each x_i is given [9].

4. CYCLEGAN

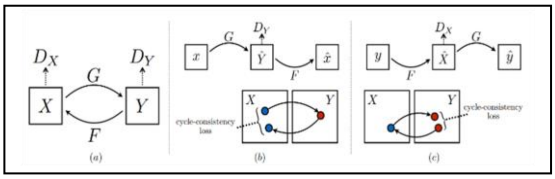

Fig. 3: The structure of CycleGan.

Fig. 3 shows the structure of CycleGan. (a) Our model contains two mapping functions G: X → Y and F: Y → X, and associated adversarial discriminators D_Y and D_X. D_Y encourages G to translate X into outputs indistinguishable from domain Y, and vice versa for D_X, F, and X. To further regularize the mappings, we introduce two “cycle consistency losses” that capture the intuition that if we translate from one domain to the other and back again, we should arrive where we started: (b) forward cycle-consistency loss: x → G(x) → F(G(x)) ≈ x, and (c) backward cycle-consistency loss: y → F(y) → G(F(y)) ≈ y. We could clearly observe that CycleGan uses the loss function to improve the quality of the learning and generation results, and to maintain the shape and structure of the input images without changing, but only the style of the images. However, in our study, not only the style changes, but also the content of the images (shape and structure of the characters) changes, so the loss function needs to be changed or selectively removed to some extent.

5. DISCOGAN

DiscoGAN [11] (Discover Cross-Domain Relations with Generative Adversarial Networks), using the discovered relations, the network successfully transfers style from one domain to another while preserving key attributes such as orientation and face identity. Moreover, pairing images can become tricky if corresponding images are missing in one domain or there are multiple best candidates. Hence, the model constructs one step further by discovering relations between two visual domains without any explicitly paired data [12].

The model for relation discovery – DiscoGAN – couples the previously proposed model. Each of the two coupled models learns the mapping from one domain to another, and also the reverse mapping to for reconstruction [13]. The two models are trained together simultaneously. The two generators G_{BA} and the two generators G_{BA} share parameters, and the generated images x_{BA} and x_{AB} are each fed into separate discriminators $L_{D_A}$and L_{D_B}, respectively.

One key difference from the previous model is that input images from both domains are reconstructed and that there are two reconstruction losses: $L_{CONST_A}$and L_{CONST_B}.

$$L_G = L_{G_{AB}} + L_{GBA} \tag{1}$$

$$= L_{GAN_B} + L_{CONST_A} + L_{GAN_A} + L_{CONST_B}$$

$$L_D = L_{D_A} + L_{D_B} \tag{2}$$

As a result of coupling two models, the total generator loss is the sum of GAN loss and reconstruction loss for each partial model (Equation 1). Similarly, the total discriminator loss L_D is a sum of discriminator loss for the two discriminators D_A and D_B, which discriminate real and fake images of domain A and domain B (Equation 2).

Now, this model is constrained by twoL_{GAN} losses and two L_{COSNT} losses. Therefore, a bijective mapping is achieved, and a one-to-one correspondence, which we defined as cross-domain relation, can be discovered.

6. EXPRIMENT

The color is a critical factor in matching the relationship between the two domains, and we also use the Otsu's method to process the anime character data.

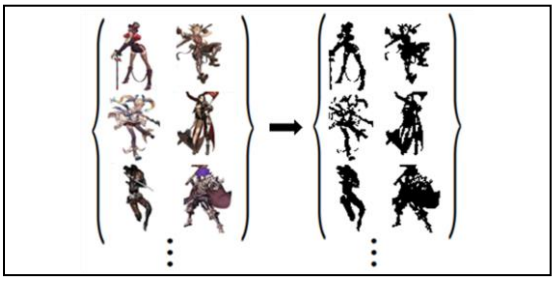

Fig. 4: Automatic image thresholding.

Fig. 4 demonstrates the results of image processing using the Otsu's method, named after Nobuyuki Otsu (Ōtsu Nobuyuki) [15], is used to perform automatic image thresholding. In the simplest form, the algorithm returns a single intensity threshold that separate pixels into two classes, foreground and background. This threshold is determined by minimizing intra-class intensity variance, or equivalently, by maximizing inter-class variance. Otsu's method is a one-dimensional discrete analog of Fisher's Discriminant Analysis, is related to Jenks optimization method, and is equivalent to a globally optimal k-means performed on the intensity histogram. The extension to multi-level thresholding was described in the original paper, and computationally efficient implementations have since been proposed.

First, CycleGAN was used for training, and the results were obtained as shown in Fig. 5. Since the output results were almost unchanged except for the pixel values, we decided to remove the consistency loss function in order to change the results.

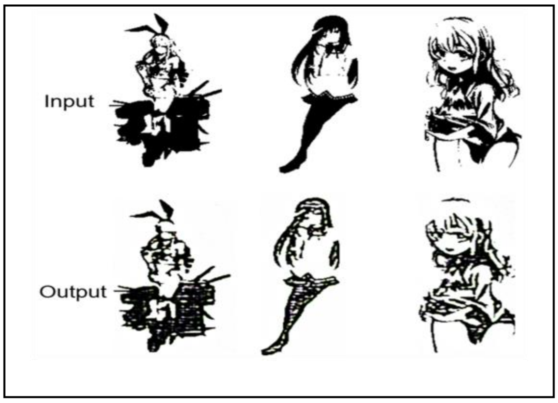

Fig. 5: Testing with CycleGAN

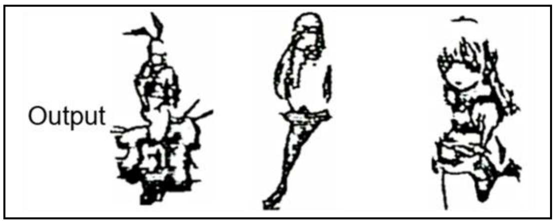

Fig. 6: Removing consistency loss

Removing the consistency loss function, the results are illustrated in Fig. 6, but still no more progress is achieved, therefore, we believe that CycleGAN is not applicable in this study. The next step will be to use DiscoGAN to re-train.

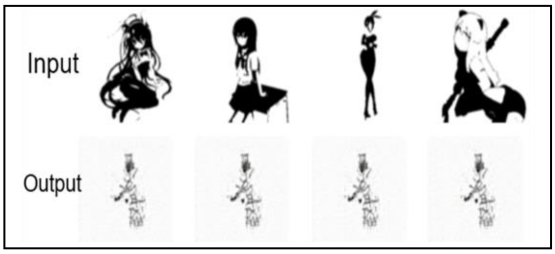

Fig. 7: Testing with DiscoGAN

We have made unprecedented progress, as shown in Fig. 7, where an image similar to the human body structure is demonstrated, although the human body structure is not clearly delineated between parts for the time being, but no longer as usual, with no change in the results.

7. CONCLUSION

We use the CycleGan model, which not only makes the style of the image slice change, but no enables the content of the image to be modified. CycleGAN is too obsessed with the retention of the content of the transformed images, and the results remain unchanged even after the removal of the consistency loss function. Compared with CycleGAN, DiscoGAN is more focused on discovering the connection between two domains. Therefore, in the subsequent research, DiscoGAN will be used for more far-reaching training, and we expect to get a better result, on which the model will be modified and adjusted to achieve an acceptable effect.

REFERENCES

[1] Larsson G, Maire M, Shakhnarovich G. Learning representations for automatic colorization[C]//European conference on computer vision. Springer, Cham, 2016: 577-593.

[2] Varga D, Szirányi T. Fully automatic image colorization based on Convolutional Neural Network[C]//2016 23rd International Conference on Pattern Recognition (ICPR). IEEE, 2016: 3691-3696.

[3] Furusawa C, Hiroshiba K, Ogaki K, et al. Comicolorization: semi-automatic manga colorization [M]//SIGGRAPH Asia 2017 Technical Briefs. 2017: 1-4.

[4] Jacob V G, Gupta S. Colorization of grayscale images and videos using a semiautomatic approach[C]//2009 16th IEEE International Conference on Image Processing (ICIP). IEEE, 2009: 1653-1656.

[5] Toshev A, Szegedy C. Deeppose: Human pose estimation via deep neural networks [C] //Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 1653-1660

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

[7] Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2223-2232.

[8] Kim T, Cha M, Kim H, et al. Learning to discover cross-domain relations with generative adversarial networks[C]//International conference on machine learning. PMLR, 2017: 1857-1865.

[9] Kotecha D. Learning cross domain relations using deep learning[D]. Dhirubhai Ambani Institute of Information and Communication Technology, 2018.

[10] Angsarawanee T, Kijsirikul B. Generating images with desired properties using the DiscoGAN model enhanced with repeated property construction[C]//Proceedings of the International Conference on Advanced Information Science and System. 2019: 1-9

[15]Otsu N. A threshold selection method from gray-level histograms[J]. IEEE transactions on systems, man, and cybernetics, 1979, 9(1): 62-66.