2022 3rd International Conference on Information Science, Parallel and Distributed Systems (ISPDS)

Generative Adversarial Networks Based on Human-Computor Interaction

2022 3rd International Conference on Information Science, Parallel and Distributed Systems (ISPDS) | 978-1-6654-8747-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISPDS56360.2022.9874027

Peiyi Jia
Master, student
School of Computer and Communication Engineering, Dalian Jiaotong University, Liao Ning
Dalian, China

Yangjie Huang
Master, student
School of Computer and Communication Engineering, Dalian Jiaotong University, Liao Ning
Dalian, China

Shijie Jia *,
Professor
School of Computer and Communication Engineering, Dalian Jiaotong University, Liao Ning
Dalian, China
jsj@djtu.edu.cn

Abstract **In order to improve the generation quality and personalization of generative adversarial network, this paper proposes an open generative adversarial network (OpenGAN) based on human-computer interaction, which adds human subjective evaluation into the training. A subjective penalty function is added to the original generator loss and the smoothing network layer is designed to reduce the impact of loss mutation in the interaction. Our results show that the IS value on ADE20K, Cityscape and other datasets increases by 61% on average, while KID and LPIPS decrease by 32% and 44%, respectively.**

Keywords-human-computer; interaction, open, generative adversarial network

I. INTRODUCTION

Generative Adversarial Network (GAN [1]) is an image generation model which have achieved the state of the art results in generating realistic images [2][4]. The key idea of GAN is to train the generator and the discriminator with alternating iterative, the discriminator adversarial loss would push the generated image to the target domain. However, traditional GAN training methods mostly have no external reference, and individuals cannot determine the quality of generated images. This paper proposes an open generative adversarial network (OpenGAN) based on human-computer interaction, which adds human subjective evaluation into the training.

The task accomplished by human-computer interaction can be regarded as a domain transformation task, that is, the process of generating a personalized high-quality image domain from the original low-quality image domain. There are two disadvantages in the current domain transformation field. First, it is easy to blur some image details to adapt to the current domain style. Second, there is still no ideal solution for the single-target transformation in the domain transformation. Human-computer interaction is a good way to solve these two problems, in which human subjective evaluations act in the training. Random score between 0 and 1 is given in each evaluation. The higher the score, the higher the quality of the generation. Evaluation results will be in the form of penalty function into the loss function, GAN will evaluate the quality of the loss function according to the penalty function, a high score will make the image continue to be generated in the same high-score way, and the loss function corresponding to low score will be punished by gradient, leading the loss function to change during the training.

The human-computer interaction proposed in this paper aims to improve the effect of domain transformation. We designed a multi-scale architecture as the human-computer interaction generator. Meanwhile, in order to verify the universality of human-computer interaction, we also added human-computer interaction to CycleGAN to compare the effects. The rest of this article is arranged as follows: the second chapter introduces the process of CycleGAN doing unsupervised tasks; the third chapter introduces the OpenGAN and loss in detail; the fourth chapter is the experimental part; finally, the thesis is summarized in the fifth chapter.

II. RALTED WORK

A. CycleGAN

CycleGAN [3] realized the unsupervised domain transformation task for the first time. It used unpaired training data to implement image domain transformation. In order to achieve this goal, it trained two GAN models simultaneously. The loss was determined by transforming the image to a combinatory mapping of a certain class. As CycleGAN needs to train two models simultaneously, it usually converges slowly, leading to long-time training process. On the basis of CycleGAN, [7] adopted one-way transformation task and used VGG as feature extractor to greatly reduce the training time. It used two encoders to encode the images separately, and then recombined the encoded results to achieve multi-modal generation. The literature [5] incorporated an attention mechanism and AdaLIN module, which improved the quality of the generation while speeding up the training.

B. Image domain transformations

Isola et al [8] proposed the first unified framework for condition-based domain transformations, which was extended

to generate high-resolution images by Wang et al [9]. Recent studies have also attempted to learn image transformations without supervision. The problem does not inherently lend itself to transformations and requires additional constraints. Some work forced the transformation to preserve certain properties of the source domain data, such as pixel values [10], pixel gradients [10], or paired sample distances. Another popular constraint is the cyclic consistency loss [3].

III. IMAGE DOMAIN TRANSFORMATION BASED ON OPENGAN

A. Interactive generative adversarial network schematic diagram

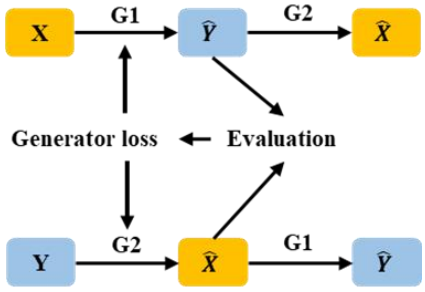

Fig. 1. Schematic diagram of human-computer interaction network generation

The human-computer interaction network schematic is shown in Fig.1. G1 and G2 are interactive domain transformation architectures. The images X and Y enter the domain transformation network G1 and G2 respectively, and output the images Yˆ and Xˆ after transformation. Then the two generated images are evaluated respectively, and the evaluation results are fed back to the domain transformation networks G1 and G2 through the generator loss function to complete the process of updating the generator parameters. we add interactive generation loss function and smoothing layer into CycleGAN, and the generators G1, G2 are designed with parameter sharing.

Human-computer interaction is mainly divided into three parts. The first part is the human-computer interaction generated network, namely G1 and G2. The second part is the image evaluation, and the third part is the loss function corresponding to the generator.

The image evaluation item is the interaction mentioned in this paper. Interaction aims at evaluating the generated image quality artificially, and feeding back the evaluation quality to the generator loss function. For high-quality images, the generator takes a low penalty and for low-quality images, the generator takes a high penalty.

B. Loss function

The loss function of the interactive generation adversarial network is:

$$L = V1(G,D) + V2(G,D) + L_{recon}^{s2} + L_{recon}^{s1} \quad (1)$$

$V1(G,D)$ is the loss function for human-computer interaction, $V2(G,D)$ is the loss function for discriminator, and L_{recon}^{s2} , L_{recon}^{s1} are the variable loss functions.

1) Generator loss function

The generator loss function in human-computer interaction is as shown in formula (2):

$$V1(G,D) = (1-\lambda) * E_{X\sim p_{data(x)}}\left[\max_{D} log\, D\left(G(x)\right)\right] + \lambda * Score * \min_{G} P(G(x)) \quad (2)$$

where '$Score$' is generally between 0 and 1through subjective evaluation. The higher the quality of the generated image, the higher the evaluation score.

$P(G(x))$ is the penalty terms, which means that when the generation quality are high, no penalty is given, and the loss function remains the same, and when the generation quality is poor, the gradient penalty is large, the network can be trained many times at this state by changing the loss function to increase the loss, so that a poorly generated image becomes a good generated image. The generator loss and the penalty loss are weighted by coefficient λ and 1-λ. The formula of penalty loss is shown in formula (3):

$$P(G(x)) = E_{X\sim p_{data(x)}}[\min_{G}\left(A_{quality}\left(G(x)\right)\right)] \quad (3)$$

2) Discriminant loss function

In order to ensure that the generated image will not produce too much blur, this paper also adds a fuzzy image set input to the discriminator loss function, which comes from the result of Gaussian filtering in X domain. The new discriminator loss function formula is as formula (4):

$$\begin{aligned} V2(G,D) = {} & E_{X\sim P(x)}[\,log\,(1-\min_{G}\max_{D} D(G(x)))] + \\ & E_{X\sim P(x)}[\max_{D} log D(x)] + \\ & E_{X\_blurry\sim P(X)}[\min_{G}\max_{D} log\,(1-D(G(X\_blr)))] \end{aligned} \quad (4)$$

X_blr is the x-domain data set after Gaussian blur, so the discriminator can optimize its minimum value to generate clearer and high-quality images.

3) Variable loss function

$$L_{recon}^{s2} = E_{c1\sim p_{data(c1)}} E_{s2\sim p_{data(s2)}} \min_{D}(G(c1,s2) - G(c1,s1)) \quad (5)$$

$$L_{recon}^{s1} = E_{c2\sim p_{data(c2)}} E_{s1\sim p_{data(s1)}} \min_{D}\left(G(c2,s1) - G(c2,s2)\right) \quad (6)$$

In the formula (5)(6), $c1$ and $s1$ represent vectors encoded by image in X-domain, $c2$ *and* $s2$represent vectors encoded by image in Y-domain, which represent the style and content of the current image, respectively. We reconstruct this part and use to generate image with content and style, and calculate the difference $G(c1,s1)$ between the original to build the minimum difference between the two contents, so as to satisfy the domain transformation with style changed and content unchanged.

C. Network smoothing layer

To verify the generation effect of human-computer interaction method, a generation model is constructed for the

experiment, as shown in Fig. 2. The generation model is constructed with convolution and residuals, and generated by multi-scale with four output results. For the horese2zebra dataset, the output image sizes are 50*50, 95*95, 180*180, 256*256, respectively. For cityscape dataset, the output image sizes are 50*100, 95*190, 180*360, 256*512, respectively. A complete generator includes G1, G2, G3 and G4 sub-generators with the same structure, At the same time, we add interactive generation loss function and smoothing layer into CycleGAN

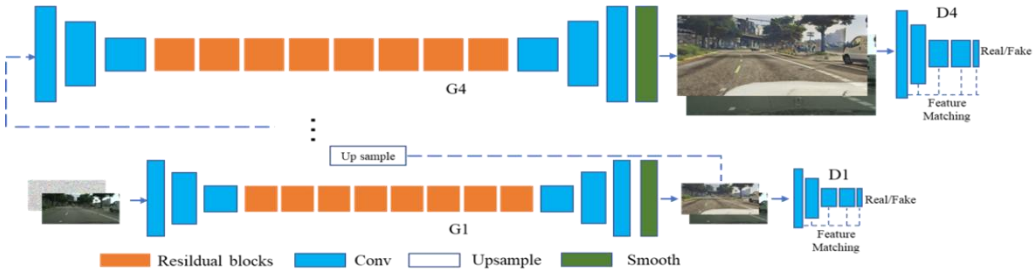

Fig. 2. Interactive generation adversarial network architecture

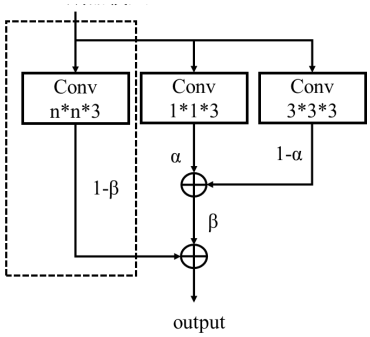

Fig. 3. Generator smoothing layer structure

to verify the universality of human-computer interaction method. In order to prevent the loss jitter in the interaction process from having a large impact on the generated image of the network, this paper adds a smoothing layer to the final part of the generator, as shown in Fig. 3. The dotted line in the figure represents the original output layer, and n is the convolution kernel size. This paper adds two parts to the output layer: the first part is 1*1*3 convolution, while the second part is 3*3*3 convolution. α, 1-α and β, 1-β all represent the output probability of the current channel. We found that the larger the receptive field features that CNN can learn, the less subtle features are needed. However, as the loss of the image will change instantaneously, we use a relatively small 3*3 convolution as a smoothing term to get better features. We set α increases with increasing number of iterations and β decreases with increasing number of iterations, as shown in formula (6).

$$\beta^{-1} \propto \sqrt{step+1}$$

$$\alpha \propto \sqrt{step+1} \tag{6}$$

IV. EXPERIMENTS

A. Settings

All algorithms in this paper are implemented in ubuntu16.04 ATL, the deep learning framework used is tensorflow2.0, the programming language is python 3.7, and the graphics card is equipped with 2*NVIDIA GTX1080Ti. The step size of each epoch is 3000, the learning rate is 0.0002, the batch size is 4, α is 0.01, and β is 0.99.

In this paper, two objective image metrics, KID and LPIPS, are used to evaluate the experimental results.

B. Experiment and result analysis

This paper compares the test results of human-computer interaction methods on three data sets, i.e. Horse2zebra, ADE20K and Cityscape.

1) Fig. 4 shows the difference between OpenGAN and CycleGAN with and without interaction

In Fig. 4, We can see that in the two sets of domain transformation results, the results generated by human-computer interaction are more realistic. Compairing the results of (c) and (e), OpenGAN is better than CycleGAN for the generation of details. Compairing (d) (e) and (b) (c), the performance of the generation network without interactive effects has a significant decline.

Since horses have fewer features than zebras, the process from horses to zebras is a conversion process from simple to complex. Under the action of human-computer interaction, poor quality images are punished and good results are obtained. In the process from zebra to horse, since it is a conversion from complex features to simple features, good results can be achieved in most domain transformation networks.

2) Fig.5 shows the comparison between CartoonGAN and OpenGAN.

The CartoonGAN and OpenGAN are trained based on the pre-training weights. From the experimental results, the artifacts generated by OpenGAN are less than that of CartoonGAN, but the quality of the domain transformation itself is not much improved. The reason is that CartoonGAN also uses the fuzzy data set as the image input, which makes the generated image with high definition.

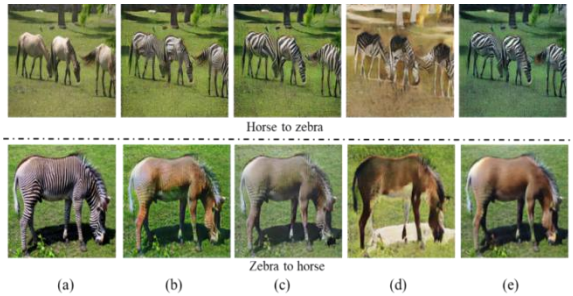

Fig.4. Comparison of the domain transformation of horse2zebra and zebra2horse. (a) is the input image. (b) is the domain transformation result of CycleGAN, (c) is the output result of CycleGAN using human-computer interaction, (d) is the output result of OpenGAN without human-computer interaction, (e) is the output result of OpenGAN.

Fig. 5 Comparison of animation style changes. (a) is the input image, (b) is the result of CartoonGAN, (c) is the result of OpenGAN.

3) Fig. 6 shows the game style comparison of U-GAT-IT, MUNIT and OpenGAN

We enlarge the details of the generation. In the transformation of game style, the color transformation is the most obvious, and the generation of texture details is also better with human-computer interaction, especially in the tree part and the house part of the images.

The LPIPS values as well as the KID/IS values are used to evaluate the image quality in Table 1. For LPIPS, a lower value means higher similarity. Considering the problem of image diversity, we do not want the difference between the result of the data set and the real image too small. For MUNIT, the multi-modal generation model should have higher image differences, but in order to compare non-multi-modal models, we have to abandon this comparison method and consider more image quality issues.

As can be seen from Table 1, the human-computer interaction method approach achieves better results in most of the datasets, but there is a significant degradation in the image generation quality without using the interaction technique, and this degradation is more pronounced for the lack of X_blr. The KID values are derived from the training weights of Inception V3. Since horse2zebra and zebra2horse are more in line with the image features of IMAGENET, the KID values of these two datasets are slightly higher than the other datasets.

TABLE1. Comparison of values under different data sets

	Horse2zebra		**Cityscape2gta**		**Cartoon style**	
	quality	**diversity**	**quality**	**diversity**	**quality**	**diversity**
Cycle-GAN	8.05/ 2.43	0.91	10.75/1.08	1.22	18.66/1.01	1.17
MUNIT	11.41/1.83	0.99	**6.88/ 1.96**	1.12	20.88/0.98	1.14
U-GAT-IT	7.06/ 2.87	1.01	9.78/ 1.12	0.97	16.21/1.12	1.01
Open-GAN	**6.81/ 2.83**	1.01	6.88/ 1.87	**1.39**	**14.69/1.14**	**1.34**
Open-GAN w/o P(G)	14.67/1.79	**1.09**	12.76/1.01	1.21	22.61/0.68	1.15
OpenGAN w/o X_blr	8.11/ 2.10	0.97	9.72/ 1.83	1.06	19.63/0.96	0.98
OpenGAN w/o interaction	22.8/ 1.01	0.66	11.86/1.06	1.39	27.66/0.71	0.76

Fig. 6. Test results from Cityscape and GTA

For the Cityscape dataset this paper retrained Inception v3 for calculating the KID values. It can be seen that OpenGAN has increased 61% in IS metrics and decreased 32% in KID.

V. CONCLUSION

We propose an open human-computer interaction method to improve the quality of image-to-image generation, which can achieve high quality and personalized generation. The results indicate that the human-computer interaction method can produce better images, but there is still a little blurring in the personalized results of the images, which is also our goal for improvement in the future. We believe that this interaction can be applied in more image fields.

REFERENCE

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A.Courville, and Y. Bengio. "Generative adversarial nets," In Conference and Workshop on Neural Information Processing Systems. (NIPS), Jun. 2014, pp. 2672–2680.

[2] R. Zhang, P. Isola, and A. A. Efros. "Colorful image colorization," In European Conference on Computer Vision. (ECCV), Sept. 2016, pp. 649-666.

[3] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros. "Unpaired image-to-image translation using cycle-consistent adversarial networks," In IEEE/CVF International Conference on Computer Vision. (ICCV), Oct. 2017, pp. 2223-2232.

[4] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. "The unreasonable effffectiveness of deep features as a perceptual metric," In IEEE/CVF Conference on Computer Vision and Pattern Recognition. (CVPR), Jun. 2018, pp. 586-595.

[5] K. Junho, K. Minjae, K. Hyeonwoo and L. Kwanghee. "U-GAT-IT: Unsupervised Generative Attentional Networks with Adaptive Layer-Instance Normalization for Image-to-Image Translation," arXiv:1907.10830v4.

[6] W. Wang, F. Zhou, W. Li, et al. "Designing the Product-Service System for Autonomous Vehicles" In IEEE IT Professional. Nov.2018, pp.62-69.

[7] X. Huang, M. Liu, S. Belongie, and J. Kautz. "Multimodal Unsupervised Image-to-Image Translation," In European Conference on Computer Vision. (ECCV), Oct. 2018, pp. 179-196.

[8] M. Mirza, S. Osindero, "Conditional generative adversarial Nets,". In Computer Science, 2014, 52(3), pp. 203-220.

[9] T. Wang, M. Liu, J. Zhu, A. Tao, J. Kautz and B. Catanzaro. "High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs," Jun. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 8798-8807.

[10] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang and R. Webb. "Learning from Simulated and Unsupervised Images through Adversarial Training," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jul. 2017, pp. 2242-2251.

[11] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, D. Krishnan. "Unsupervised Pixel-Level Domain Adaptation with Generative Adversarial Networks." 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jul. 2017, pp. 95-104.