Cycle-Consistent Adversarial Network for Facial Local-Region Exchange in Wild

Xiao Sun, Pingping Xia and Fang Tian

May 26, 2019

Cycle-Consistent Adversarial Network for Facial Local-Region Exchange in Wild

1st Xiao Sun
School of Computer and Information
Hefei University of Technology
Anhui, China
sunx@hfut.edu.cn

2nd Pingping Xia
School of Computer and Information
Hefei University of Technology
Anhui, China
2017111015@mail.hfut.edu.cn

3nd Fang Tian
Modern Education Technology Center
Qinghai University
Qinghai, China
tianfang@qhu.edu.cn

Abstract—Recently, Generative Adversarial Networks are popularly used in face generation and get the state-of-art result. However, it's hard to swap the local area of face while many of previous work has focused on either generating face from a noise vector which belongs to some kind of data distribution or swaps the whole face. In this paper, we proposed a Cycle-Consistent Region Exchange Generative Adversarial Network(CREGAN) for facial local area exchange in the wild facial database. The Cycle-Consistent guaranteed that the exchanged area keeps the another facial feature and a novel approach to achieve face local region exchange and other region remain unchanged. At the same time, the characteristics of generative adversarial network can make ensure the quality of the generated images. And, it will shows that the generated images can reach photo-realistic results by CREGAN.

Index Terms—Generative Adversarial Network, Cycle-Consistent, Wild Facial Database

I. Introduction

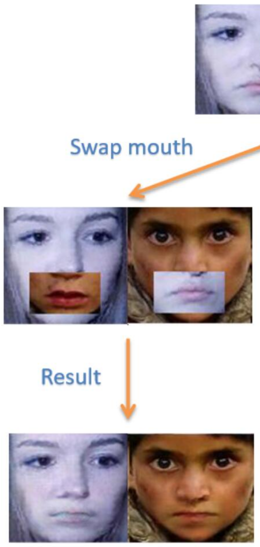

Fig. 1: Local region exchange instance by the CREGAN.

Many models based on Generative Adversarial Networks(GANs) [1] have been proposed for face generation by scholars and research enthusiasts, such as [2]–[6], and have obtained the remarkable results. EBGAN [2] and BEGAN [3] have been proposed to generate a face image from a random noise vector which belongs to a certain distribution. However, it is uncontrollable that the content generated by this way, that is, although a face can be generated, it's hard to control whether it is a man or a woman. There is another way to locally generate a face by image completion [4], [5], which has been used for image repair. And it still can't control the changed area with specific features. Of course, there are also some content-based GAN models, such as Conditional GAN [7] and InfoGAN [8], which used the attribute tags to guide the generator to generate face images with specific content from a noise vector. But these tasks are difficult to generate local regions with specific tags. There is also same for the whole face exchanges [6] to only exchange local part of a face.

We proposed the CREGAN which was inspired by image-to-image translation [9], [10], the whole face swap [6], and image completion [4], [5], [11] to achieve two main goals. One is that it can use the local area of the first face to replace the relative area of the second face between two face images, thus forming a new face image, and vice versa. Another is to make sure the area where the replaced area retains the feature of another image, while the area non-replaced remains unchanged. The Figure 1 shows what have done.

In view of the above two goals, there are mainly two difficulties, which will further discuss in Section III. The first difficulty is how to ensure that the replaced area and the non-replaced area to retain the original features. And another is how to ensure that the replaced area is combined with the non-replaced area to form a photo-realistic face image.

In ordered to solved the problems above, we designed the CREGAN base the Generative Adversarial Networks [1] which has achieved remarkable results in the field of image generation. But using DCGAN [12] to train our work, it's hard to generate photo-realistic image, especially in the boundary of two area fusion. Inspired by [4], [5], the CREGAN adapted a global and local discriminator module to generate photo-realistic result can work with local area of any size.

Finally, the CREGAN has been used to exchange the local regions in different faces. In this paper, the eye and mouth regions have been selected to exchange to verify the effect of the CREGAN, because of their rich characteristics of a face.

Overall, our contributions are as follows:

A Cycle-Consistent loss has been designed for Generator to generate image, which guarantees replaced region with another image features and non-replaced region remain unchanged in CREGAN.

A Global and Local Discriminator has been proposed for CREGAN to identify the true and false samples by global and local input image.

Finally, the CREGAN has been used for face local part exchange and get photo-realistic results.

II. Related Work

GANs[3] quote the idea of minimax game, leading to a Generator and a Discriminator. The goal of the Generator is to generate images to confuse the Discriminator by learning real data distribution. At the same time, the purpose of the Discriminator is to determine whether an image is generated or real. GANs have been widely used in the field of image generation to, such as face generation [2], [3], [13]–[15], image-to-image translation [9], [10], image completion [4], [5], [11], etc. BEGAN [3], was proposed to generate facial image from a random noise vector which belonged to a kind of distribution. Cycle-GAN [9] made use of a Cycle-Consistent loss to realize style transition between two kinds of different style pictures, such as zebras trans to hoses or horse trans to zebras. But it's can only deal with two kinds of different categories of image transformation. Then, StarGAN [10] was proposed to achieve multi-domain image-to-image translation, which adopted the idea of Cycle-Consistent in [9]. In this paper, inspired by [9], [10], a Cycle-Consistent loss has been presented for face part exchange which ensures that the replace area and non-replaced area remain features consistent.

Global and Local Discriminator has been mainly used in image completion [4], [5] to jointly judge the authenticity of the images from global area and local area. However, when processing the local area, they all need to pick it out, and resize it to a fixed size for Discriminator. In dealing local area with larger length-width ratio or local area size is too big or too small, it's easy to lost some details characteristics. Thus we put forward a Global and Local Discriminator can work with local area of any size.

However, when training GAN, there is a problem about the stability and convergence of the model are difficult to be guaranteed. Much work has been done to solve this problem [12], [16]–[18]. Where, Wasserstein GAN(WGAN) [17] analyzed the problem in general theory ,and gave the improvement skills through formula reasoning. WGAN-GP [16] is an improvement for WGAN which accelerated the convergence speed of the model by introducing gradient penalty instead of Lipschitz constraint. In this paper, we also used the WGAN-GP to train the CREGAN.

III. Proposed method

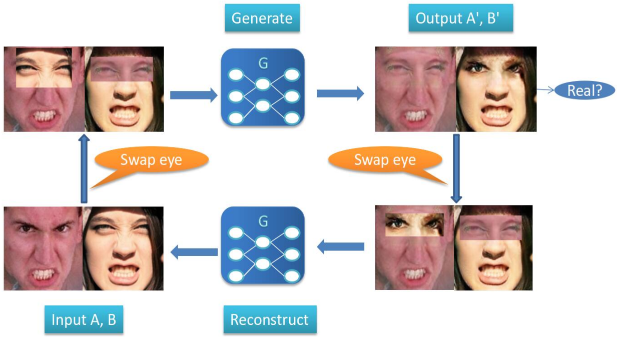

Fig. 2: Generator Network.

In order to describe the CREGAN in detail, the symbol representations are defined as follows. Training image set: X=$[X_1, X_2, ..., X_i, ..., X_N]$, $X_i \in R^{H*W}$, where N is the number of training set, H and W is the hight and width of one image. Exchange area mask: EM=$[EM_1$, EM_2,...,EM_i,...,$EM_N]$, $EM_i \in R^{H*W}$, where the value of exchange area is zero, other value is one and it should be noted that the EM_i is square area while top-left point is (a_1, b_1) and bottom-right point is (a_2, b_2). The mapping of Generator is denoted as G(·) and the Discriminator is D(·)

A. Generator

The purpose of the generator is to exchange the corresponding area between two face images with the area exchanged in the generated image retains other's feature, while the area not exchanged in the generated image keeps the features of their own. Therefore, it can make up the original image by transplanting the replaced area of the generated image to the first image, or vice versa.

The Generator network structure is shown in Figure 2. In Figure 2, given two input A, B, the Generator takes the images after their eye area is exchanged as input and generates fake images A' and B' which are distinguished by Discriminator. Then exchanging the swap area between fake images, and input the result to the Reconstruct to generate the A and B. The parameters of Generator and the Reconstruct are shared.

Selecting any two face image (X_i, X_j) front the training set, and the exchanged area mask is (EM_i, EM_j). It is important to note that (EM_i, EM_j) represent the same replacement area, such as the mouth or eyes. For the sake of simplicity, combining (EM_i, EM_j) in one EM_{ij} by max operation.

Then get the input X_i' and X_j' of G(·) as follows:

$$\begin{aligned} X_i' &= X_i * EM_{ij} + X_j * (1 - EM_{ij}) \\ X_j' &= X_i * (1 - EM_{ij}) + X_j * EM_{ij} \end{aligned} \tag{1}$$

where * is corresponding element multiplication. Denote X' as the input of the G(·).

So, the input of Reconstruct operation can be represented as:

$$\begin{aligned}\overline{X}_i &= G(X_i') * EM_{ij} + G(X_j') * (1 - EM_{ij}) \\ \overline{X}_j &= G(X_i') * (1 - EM_{ij}) + G(X_j') * EM_{ij}\end{aligned} \tag{2}$$

Therefore, a Cycle-Consistent loss has been proposed to ensure the goal of the exchanged area retains the features of the another image and the other area remain unchanged. The loss as follows:

$$L_{rec} = ||G(\overline{X}_i) - X_i||_1 + |||G(\overline{X}_j) - X_j||_1 \tag{3}$$

where the L1 norm [19] has been adopted as the reconstruct loss.

It used the dilated convolutional layers [20] to increase the receptive field of the convolution operation to make the replaced region refer to a broader regional characteristics, mainly color features in the generator network.

The specific architecture of the network is shown in Table I, which Conv is the standard convolution operation and Dila-Conv is the dilated convolution operation. Each Operation is followed by a instance normalization [21] and Leaky ReLU [22] except the last Conv which is followed by Tanh activation.

TABLE I: Gnerator network architecture

Operation	Kernel	Dilation	Stride	Output
Conv	7*7	1	1*1	64
Conv	4*4	1	2*2	128
Conv	4*4	1	2*2	256
Conv	3*3	1	1*1	256
Dila-Conv	3*3	2	1*1	256
Dila-Conv	3*3	4	1*1	256
Dila-Conv	3*3	8	1*1	256
Dila-Conv	3*3	16	1*1	256
Conv	3*3	1	1*1	256
Conv	3*3	1	1*1	256
De-Conv	3*3	1	2*2	128
De-Conv	4*4	1	2*2	64
Conv	4*4	1	1*1	3

B. Discriminator

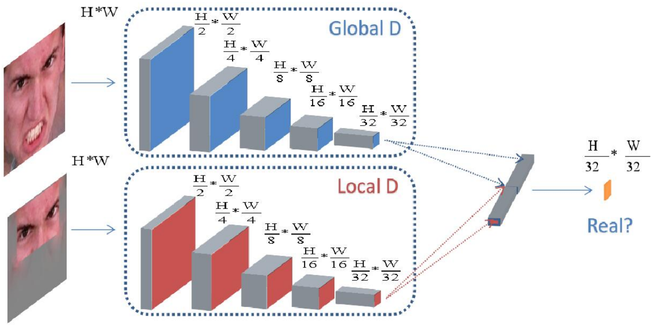
H*W
$\frac{H}{2}*\frac{W}{2}$
$\frac{H}{4}*\frac{W}{4}$
$\frac{H}{8}*\frac{W}{8}$
$\frac{H}{16}*\frac{W}{16}$
$\frac{H}{32}*\frac{W}{32}$
Global D
H*W
$\frac{H}{2}*\frac{W}{2}$
$\frac{H}{4}*\frac{W}{4}$
$\frac{H}{8}*\frac{W}{8}$
$\frac{H}{16}*\frac{W}{16}$
$\frac{H}{32}*\frac{W}{32}$
Local D
$\frac{H}{32}*\frac{W}{32}$
Real?

Fig. 3: Discriminator network.

The main purpose of the Discriminator is to determine whether an image is generated or real. As demonstrated in Figure 3, a Global and Local Discriminator has been proposed to judge the authenticity of images by combing the global and local features. Unlike [4], it do not need to pick local area out, and interpolate it to a fixed size for Discriminator. Here a distance loss function has been used to generate a local mask to process the image as input to the D(·).

Given the the input X_i of D(·) and the exchanged area mask EM_i, it can get exchanged area top-left points(a_1, b_1) and bottom-right area points(a_2, b_2). Local Mask: LM_i as follows:

$$LM_i[m,n] = \begin{cases} 1, if: a_1 \leq \ m \leq \ a_2, b_1 \leq \ n \leq \ b_x \\ \gamma^{kd}, else. \end{cases} \tag{4}$$

where $d = min(m-a_1, a_2-m, n-b_1, b_2-n)$, k is a distance attenuation factor, γ is a base of exponential function.

Thus, there are two input for D(·), the global image and the global image with local mask. The parameter of D(·) is shown in Table II. Each operation is followed by a instance normalization and Leaky ReLU except the Concatenate Layer without any activation.

TABLE II: Discriminator

(a) Global Discriminator

Operation	Kernel	Stride	Output
Conv	4*4	2*2	64
Conv	4*4	2*2	128
Conv	4*4	2*2	256
Conv	4*4	2*2	512
Conv	4*4	2*2	1024
Conv	4*4	2*2	1024

(b) local Discriminator

Operation	Kernel	Stride	Output
Conv	4*4	2*2	64
Conv	4*4	2*2	128
Conv	4*4	2*2	256
Conv	4*4	2*2	512
Conv	4*4	2*2	1024
Conv	4*4	2*2	1024

(c) Concatenate Layer

Operation	Kernel	Stride	Output
Concat	-	-	2048
Conv	3*3	1*1	1

Finally, the objective functions of the CREGAN can be written as:

$$\begin{aligned} L_D = E[D(X, LM)] - E[D(G(X'), LM)] \\ + \lambda_{gp} E[(||\Delta; D(X, LM)||_2 - 1)^2] \end{aligned} \tag{5}$$

$$L_G = E[D(G(X'), LM)] \tag{6}$$

where λ_{gp} is gradient penalty factor.

IV. Experiments

This section will introduce the facial databases, the training process of the CREGAN in detail, and some settings of hyper-parameter.

A. Database

RAF Database [23], is a large-scale facial expression database with about 15K great-diverse facial images which train set is 12K and test set is 3K in 7 classes of basic emotion such as anger, disgust, fear, happiness, neutral, sadness, surprise.

Affect Database [24], is a wild data sets on facial expressions, valence, and arousal.This data sets contains about 410 k manually tag images and about 550 k automatic marking image by ResNext Neural Network training on manually tag data model to tag.

B. GAN training

Before training, we should to define the value of some hyper-parameters, γ=0.9 , k=1 in Eq. (4). The optimizer algorithm used if the Adam [25] optimizer with β_1=0.5, β_2=0.9, and learning rate is 0.001 which is reduced by base 0.9 every 100 iterations. The training process is Algorithm 1, which the mini-batch size is 16, T_{train} is 1000, $train_G$ is 3.

V. Conclusion

The CREGAN has been proposed in this paper, a novel approach to achieve facial local area exchange. In order to achieve the above objectives, a Cycle-Consistent loss has been presented to guarantee the exchanged region with another image features and the other area remain unchanged. In the meanwhile, in order to ensure that the resolution of the exchanged image can reach photo-realistic, the idea of GANs have been adopted. A Global Local Discriminator has been proposed to learn how to fuse the features like the borders and colors between exchanged area and the other area. Finally, the CREGAN has been used for face local part exchange and get photo-realistic results.

Acknowledgment

This work was supported by the State Key Program of the National Natural Science Foundation of China (61432004, 71571058, and 61461045). This work was partially supported by a project funded by the China Postdoctoral Science Foundation (2017T100447). This research was also partially supported by the National Natural Science Foundation of China under Grant No. 61472117, by a Qinghai Province Science and Technology Fund for fundamental and applied research (No. 2016-ZJ-743), and by the Open Project Program of the National Laboratory of Pattern Recognition (NLPR).

Algorithm 1 CREGAN training procedure

```
1:  while iterations $t < T_{Train}$ do
2:      # Exchange mouth area
3:      Sample a mini-batch of images $X_1$ and $EM_1$(mouth exchanged mask). Thereby get $LM_1$(the local mask of D(.) in Eq. (4));
4:      And Random $X_1$ permutation get $X_2$ , $EM_2$ , and $LM_2$;
5:      For mouth, it can get $X_1'$ and $X_2'$ by Eq. (1) for G(.), and the real input $[X_1/LM_1, X_2/LM_2]$, the fake input $[G(X_1')/LM_1, G(X_2')/LM_2]$ for D(.). So the eye is same.
6:      Update D(.) by $LD$ by Eq. (5);
7:      if t% $train_G$ ==0 then
8:          Update G(.) by $LG$ by Eq. (6);
9:      end if
10:     # Exchanged eye area
11:     Repeat 3-9 while $EM_1$ is eye exchanged mask;
12: end while
```

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Advances in Neural Information Processing Systems 27, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2672–2680.

[2] J. J. Zhao, M. Mathieu, and Y. LeCun, "Energy-based generative adversarial network," CoRR, vol. abs/1609.03126, 2016.

[3] D. Berthelot, T. Schumm, and L. Metz, "BEGAN: boundary equilibrium generative adversarial networks," CoRR, vol. abs/1703.10717, 2017.

[4] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," ACM Transactions on Graphics (Proc. of SIGGRAPH), vol. 36, no. 4, pp. 107:1–107:14, 2017.

[5] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," arXiv preprint arXiv:1801.07892, 2018.

[6] H. Dong, P. Neekhara, C. Wu, and Y. Guo, "Unsupervised image-to-image translation with generative adversarial networks," CoRR, vol. abs/1701.02676, 2017.

[7] M. Mirza and S. Osindero, "Conditional generative adversarial nets," CoRR, vol. abs/1411.1784, 2014.

[8] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in NIPS, 2016.

[9] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networkss," in Computer Vision (ICCV), 2017 IEEE International Conference on, 2017.

[10] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," arXiv preprint arXiv:1711.09020, 2017.

[11] B. Dolhansky and C. Canton-Ferrer, "Eye in-painting with exemplar generative adversarial networks," CoRR, vol. abs/1712.03999, 2017.

[12] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," CoRR, vol. abs/1511.06434, 2015.

[13] R. Huang, S. Zhang, T. Li, and R. He, "Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis," in The IEEE International Conference on Computer Vision (ICCV), Oct 2017.

[14] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in Advances in Neural Information Processing Systems 29, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Curran Associates, Inc., 2016, pp. 469–477.

[15] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," 10 2017.

[16] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," CoRR, vol. abs/1704.00028, 2017.

[17] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," CoRR, vol. abs/1701.07875, 2017.

[18] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in NIPS, 2016.

[19] E. W. Weisstein, "L1-norm," in From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/L1-Norm.html.

[20] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in ICLR, 2016.

[21] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," CoRR, vol. abs/1607.08022, 2016.

[22] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in in ICML Workshop on Deep Learning for Audio, Speech and Language Processing, 2013.

[23] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017, pp. 2584–2593.

[24] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," IEEE Transactions on Affective Computing, vol. PP, no. 99, pp. 1–1, 2017.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," CoRR, vol. abs/1412.6980, 2014.