Face Photo-Sketch Synthesis Based on Conditional Adversarial Networks

Xiaoqi Xie[1], Huarong Xu[2], and Lifen Weng[3(✉)]

[1] Electrical Engineering, Xiamen University of Technology, Xiamen, China
xqixie@163.com
[2] Computer Science and Technology, Xiamen University of Technology, Xiamen, China
hrxu@xmut.edu.cn
[3] Academy of Design Arts, Xiamen University of Technology, Xiamen, China
2009990509@xmut.edu.cn

Abstract. Face Photo-Sketch synthesis is designed to generate an image contains rich personal information and details facial. It is widely used in law enforcement and life areas. Although some existing methods have achieved remarkable results, however, due to the gap between the synthetic and real image distribution, the synthetic image does not achieve the expected effect. To solve this problem, In this paper, we proposed conditional generation adversarial networks (CGAN) based on arts drawing steps constrain. We divide the process into two stages. In the first stage, we take the original face photos which concat noise z as input, generating stage 1 low resolution synthesize sketches. In the second stage take the stage 1 results and original face photo as inputs, yielding stage 2 high resolution synthesize sketches, which can express more natural textures and details. To add realism, we train our network using an adversarial loss. Experiments have shown that, Compared with previous methods, our results generate visually comfortable face sketches. And express more natural textures and details.

Keywords: Face Photo-Sketch synthesis ·
Arts drawing steps constrain ·
Conditional generation adversarial networks

1 Introduction

The face sketch has rich shadow texture and strong three-dimensional texture, which can vividly express the characteristics and personality of the character, so it is widely used in law implementation. For example, in the process of law enforcement, in most cases, the real face of the suspect cannot be obtained. At this time, the best alternative is usually based on sketches of eyewitness recalls which can help the police quickly narrow down and lock the suspect [5,11,13,18,25]. In terms of life crafts, passenger often keep a sketch by the painter in a tourist attraction or scenic spot as a souvenir. In addition, it has

Supported by Project of Fujian Science and Technology Department 2019I0036.

W. Ni et al. (Eds.): WISA 2019, LNCS 11817, pp. 59–71, 2019.
https://doi.org/10.1007/978-3-030-30952-7_7

a wide range of applications, such as digital entertainment [9,14]. At present, face sketches are mostly drawn by painters or professional artists create them through painting tools. They need long-term practice to have this skill, and this kind of creative method is inefficient. So, the automatic generated face sketch technique is becoming the research hot in these application area.

There are some algorithms such as Sketch2Photo [4] and Photos-ketcher [1], require a large number of fine extraction features, and need complex processing as cutting images to make the image more realistic. In recent years, the convolutional neural network in deep learning is very popular and provides a powerful method for facial sketch synthesis [3,19,31]. Among them, the extended GAN network [8] application to the face sketch synthesis method is especially prominent. Image-to-Image Translation with Conditional Adversarial Networks [8] uses CGAN to solve face sketch synthesis. This network can not only learn the mapping relations between input image and output image, but also learn the loss function for training mapping relations. Unpaired Image-to-Image Translation Zhu et al. [35] proposed use Cycle-Consistent Adversarial Networks (CycleGAN) to solve where paired training data does not exist problem. The article introduces a cycle consistency loss to make $F(G(X)) \approx X$ (and vice versa). However, the meaning is to convert source domain images back into target domain images, and its distribution is still the same.

Although these methods have achieved remarkable results, due to the limitations of their structural consistency, accurate depicting face photos/sketches remains challenging. In addition, the synthesized image mostly yield blurred effects, leading to the synthesized image is unrealistic. In order to solve this problem, we proposed conditional generation adversarial networks (CGAN) based on arts drawing steps constrain.

2 Related Work

In generation adversarial network we must train two models at the same time. The generator mainly learns the real image distribution to make the image generated by itself more realistic, to fool the discriminator. The discriminator needs to make a true and false judgement on the received picture. The probability that the prediction of the fixed image is true is close to 0.5 [6]. Since then, many improvements and interesting applications have been proposed [7]. recently, the great success achieved by conditional Generative Network(cGAN) [15]. In a series of image-to-image translation tasks [10]. For example, CycleGAN [35] generate high quality sketches using multi-scale discriminators [26]. Wang et al. [27] recommend to first generate a sketch using the vanilla cGAN [26] and refine it using a postprocessing method called back projection. Jun Yu et al. [12] proposed to improve the CA-GAN after GAN to generate a realistic synthesized sketch. Di et al. combining a convolutional variational Autoencoder with cGAN for attribute-aware facial sketch synthesis [3]. Stacked networks have also made great progress in this area, such as image super-resolution reconstruction and generate high resolution image with photo-realistic details [12]. They have all

achieved very good results. Affected by these success stories, we are interested in using CGAN to generate sketch photos, but found that the effect is not the best, so it is proposed based on arts drawing steps constrain CGAN to improve the effect. Our method divide into two stages and is different from the previous one when choosing generation and discriminator. We are Using Encoder-Decoder [2] as Generator, and for discriminator we use patchGAN [16].

Encoder-Decoder is a very common model framework for deep learning. Many previous solutions [10,29,30,34] to problems in this area have used an encoder-decoder network [17]. For example Unsupervised algorithm auto-encoding [2], application of image caption, the neural network machine translation NMT model all trained with encoder-decoder. Therefore, encoder-decoder not a model, but a type of framework. The Encoder and Decoder sections can be anything, as text, voice, image. Models can be CNN [32], RNN, LSTM, and more. So we choose Encoder-Decoder. It is an end-to-end learning algorithm.

GANs network structure has been shown in some experiments to be unsuitable for the field of images requiring high resolution and hight detail retention. Some people have designed PatGAN [16] based on this situation.The difference between PatchGAN is in the discriminator. The general GAN is a vector that only need to output a true or false which represents the evaluation of the whole image, but the PatvhGAN output is an $N \times N$ matrix, each element such as a(i,j), has only two choices of True or False(label is a matrix of $N \times N$, each element is True or False), and the result is often through the convolution layer to achieve. The discriminator makes a true and false discriminant for each patch, and average the results of all patches of a picture as the final discriminator output.

3 Methodology

3.1 Model

Figure 1 shows the model structure. We all know in order to express details and textures, the painter's painting is divided into many steps. In order to produce high resolution sketches like a painter, we divided our process into two stages based on arts drawing steps constrain. In the first stage, we take the original face photos which concat noise z as input, generating stage 1 low resolution synthesize sketches x_{g1}. In the second stage take the stage 1 results x_{g1} and original face photo as inputs, yielding stage 2 high resolution synthesize sketches x_{g2}, which can express more natural textures and details. G we choose encoder-decoder, D we use PatchGAN.

Generator(G1,G2). The G1 network take photos and noise as input, G1 structure as follows: Conv1(16)-Conv2(32)-Pooling2(32)-Conv3(64)-Poling 3(32)-Conv4(128)-Conv5(128)-Deconv6(128)-Deconv7(64)-conv8(32)-conv9(1). Conv(K) denotes K-channel convolutional layers, pooling is max-pooling. Convolution and polling layers are used to extract advanced feature. G2 network have two modules. The input one is G1 generated synthetic sketches and another input is photos. G2 structure same as G1.

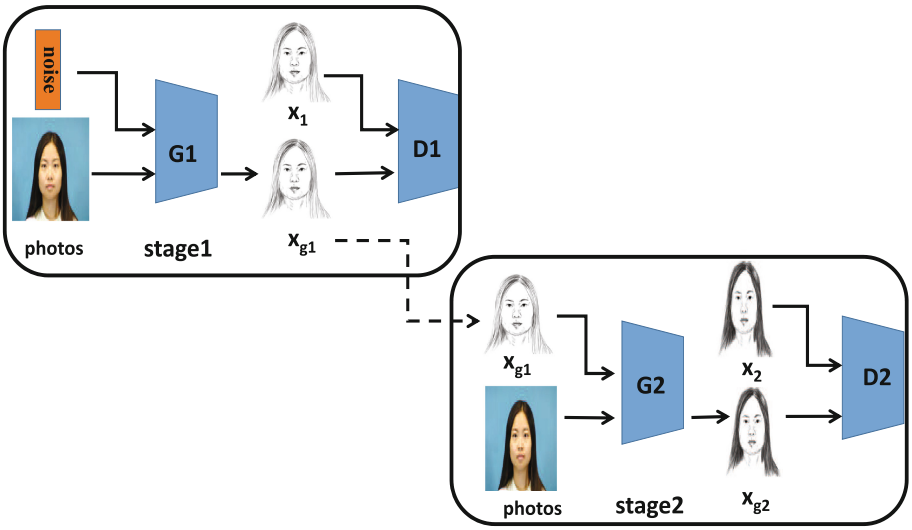

Fig. 1. Two-stage trainging networks. Stage 1 generates x_{1g}. Stage 2 concat photos and x_{1g} put into G2 generates sketch image.

Discriminator(D1,D2) using patch-Decoder discriminator D1, there are 16 patches in every iterative training with G1. PatchGAN discriminator preserving of high-frequency details. All the convolutional layers in D1 have a filter size of 3×3. D2 structure same as D1.

3.2 Data Set

We build a two stages face photo-sketch database. 188 faces provided by CUHK students database [24]. For each face, there are two stages sketches by our professional painter and a photo taken in a frontal pose, under normal lighting condition. In CHUK student database, we choose 88 faces for training and 100 faces for test.

3.3 Objective Function

The GANs there are two networks generators and discriminators. The goal of G is to generate a real images as much as possible to deceive the discriminant network D. The goal of D is to separate the G generated image from the real image. Conditional GAN is another variant where the generator is conditioned on additional variables such as category labels, partial data for image restoration, date from different modalities. If the conditional variables is a category label, it can be seen that CGAN is an improvement to turning a purely unsupervised GAN into a supervised model. The objective of a conditional GAN can be defined as:

$$L_{cGAN}(G, D) = E_{x,y}[logD(x, y)] + E_{x,z}[log(1 - D(x, G(x, z)))] \quad (1)$$

where z, is noise as input, x, is a corresponding face image, y, the output images, are sampled from the true data distribution $P_{data}(x, y)$.

To add realism, we train our network using an adversarial loss, similar to Generative Adversarial Networks(GANs) [6], such that the synthesized sketches are indistinguishable from target photos. In order to measures the error between the synthesized sketch image and target sketch, we complement the adversarial loss with a regularization loss [8,21,23]. Stage 2 is built on stage 1 to generate realistic target sketches. where $x_g = (x, z)$ is generated by stage 1. Stage 2 not using noise, assuming that the noise has already been maintained in the stage 1 generated x_g, Our final objective is define as follow:

$$G^* = L_A + \lambda L_1 \tag{2}$$

where L_A is the adversarial loss, L_1 is the loss based on the L_1-norm between the synthesized image and the target. λ is weight. G is tries to minimized this loss to updates its parameters, against D is that tries to maximized loss is defined as

$$L_A = \min_G \max_D L_{cGAN} \tag{3}$$

The L_1 loss is present synthesized sketch image and the target sketch loss. x indicate target sketch, x_g indicate synthesized sketch.

$$L_1 = \| x - x_g \|_1 \tag{4}$$

3.4 Optimization

To optimize our networks, follow [8,12] we alternate between one gradient descent step on D, then one step on G. All networks are trained using Adam solver. In our experiments, we use batch size 1 and momentum parameters $\beta = 0.9$. Follow the GAN approach [14], we model include two-stage, every stage has a generator and a discriminator. which are sequentially denoted by $G1$, $D1$, $G2$, $D2$. As follows Algorithm 1.

4 Experimental Results

4.1 Data Set

We build a two stages face photo-sketch database. 188 faces provided by CUHK students database [24]. For each face, there are two stages sketches by our professional painter and a photo taken in a frontal pose, under normal lighting condition. In CHUK student database, we choose 88 faces for training and 100 faces for test.

In Sect. 4.2, we use CUHK database train our model and do three experiments. first, experimental setting and evaluation of the proposed method are discussed in detail. Results are compared with previous state-of-the-art-methods. Second, we compare with who also use two-stage generation model. Third, we

do internal comparisons, using the same structural model for one-stage and two-stage sketch generation comparisons. We then provide a quantitative comparison with the method mentioned in the first part in Sect. 4.2. Finally, we show a few examples of our model results in Fig. 5.

Alogrithm 1 Optimization procedure of our network

Input: Set of synthetic images x_{g1} ,x_{g2} ,and real images x_1,x_2,a face photo y,max number of steps(T);

Outoput: optimal G1,D1,G2,D2;

initial G1,D1,G2,D2;

for t=1,. . .,T **do**

1. select one training instance:

{Set of synthetic images x_{g1} ,x_{g2} ,and real images x_1,x_2,a face photos y}

2. Estimate the first stage sketch imag:

x_{g1} =G1(y,z)

3. Estimate the first stage sketch imag:

x_{g2} =G2(x_{g1} ,y)

4.Update D1:

D1=argmin$_{D1}L_A$(G1,D1)

5.Update D2:

D2=argmin$_{D2}L_A$(G2,D2)

6.Update G1:

G1=argmax$_{G1}L_A$(G1,D1)+$\lambda L_1(x_1,x_{g1})$

7.Update G2:

G2=argmax$_{G2}L_A$(G2,D2)+$\lambda L_1(x_2,x_{g2})$

end

4.2 Qualitative Results

In the first set of experiments, we perform face photo-sketch synthesis on the CUFS datasets and compared with existing advanced methods: MWF [33], MRF [28] and cGAN [8]. And we attempt to recover the image directly from attributes without going to the intermediate stage of sketch. The entire network in Fig. 2 is trained stage-by-stage using caffe.

In the second set of experiments. As show Fig. 3, we compare results with CA-GAN and SCA-GAN [12], which also uses a two-stage generation model.

In the third set of experiments, And we attempt to recover the image directly without going to the intermediate stage of sketch. As show in Fig. 4. And we show some our model results in Fig. 5.

4.3 Quantitative Results

In the CGNs, the discriminator and generator objective functions are usually used to measure how they do each other. But this does not measure the quality and diversity of the generated images. Usually, we use the IS (inception score) and FID (Fr'echet Inception distance) indicator to evaluate different and GAN model. In this work, we choose the Fr'echet Inception distance (FID) to evaluate

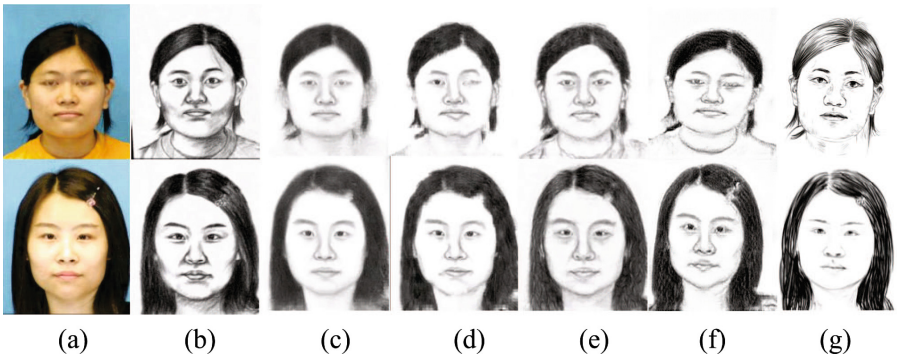

Fig. 2. Comparison of the sketch synthesis result using three different approach and our model: (a) photos; (b) sketches by artist from CUFS; (c) MWF; (d) MRF; (e) cGAN; (f) our model; (g) sketches by our artist

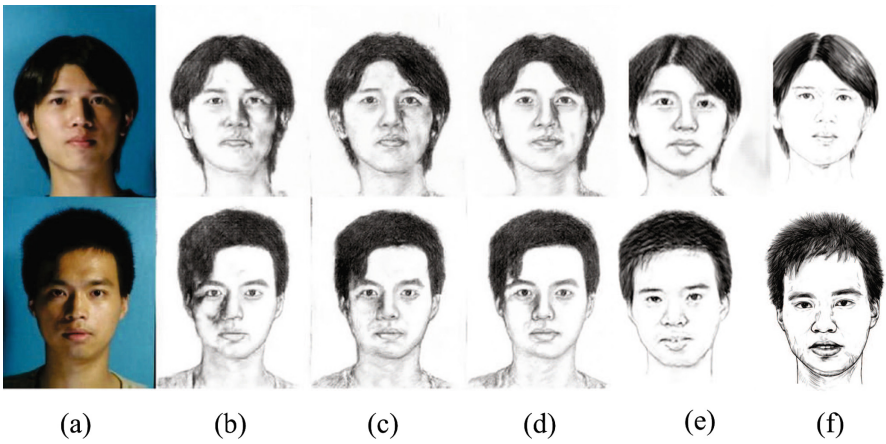

Fig. 3. (a) photos; (b) cGAN; (c) CA-GAN; (d) SCA-GAN; (e) our model; (f) sketches by our artist

the realism and variation of synthesized sketches [12,20]. FID values present distances between synthetic and real data distributions. Therefore, lower FID is better. We use all test samples to compute the FID. In FID we use inception network to extract the characteristics of the middle layer [12,22].

We also use the Feature Similarity index Metric (FSIM) [22] between the synthesized image and the corresponding ground truth image to objectively evaluate the quality of the synthesized image [12]. These two indicators have been compared in article [12], we compare the experimental data of our model based on this article. FID the smaller the better. FSIM the bigger the better. The comparison results are shown in Table 1. And we comparison results with one-stage synthesized sketch In Table 2, which show two-stage synthesized advantage.

Here we have to declare that our database is also a CUFS database, but we use a two-stage sketch drawn by our artist, which is different from the CUFS sketch used by other methods. So the quantitative analysis given here is also relatively comparative. In a sense, it has reference value.

The results are presented in Table 1. The proposed two-stage method produces the relatively low FID. Implying that our method generated images are more realistic than the ones generated by most previous methods. our method produces the higher FISM. In Table 2, compared one-stage and two-stage, two-stage achieves lower FID values and higher FSIM vales than one-stage.

Table 1. Quantitative results corresponding to different methods use CUHK dataset.

Criterion	dataset	MWF	MRF	cGAN	CA-GAN	SCA-GAN	Our model
FID	CUHK	87.0	68.2	43.2	39.7	36.2	34.4
FSIM	CUHK	71.4	70.4	71.1	71.2	71.4	72.7

Table 2. Quantitative results corresponding to our methods one-stage and two-stage use CUHK dataset. One-stage: Generate directly in one step, input the original image output sketch. In order to highlight the advantage of the two-stage generation model.

Criterion	dataset	one-stage	two-stage
FID	CUHK	44.3	34.4
FSIM	CUHK	69.8	72.7

Fig. 4. (a) photos; (b) one-Stage results. Omitting a stage, directly using the results of one stage test; (c) two-Stage results(output from our method); (d) sketches by our artist

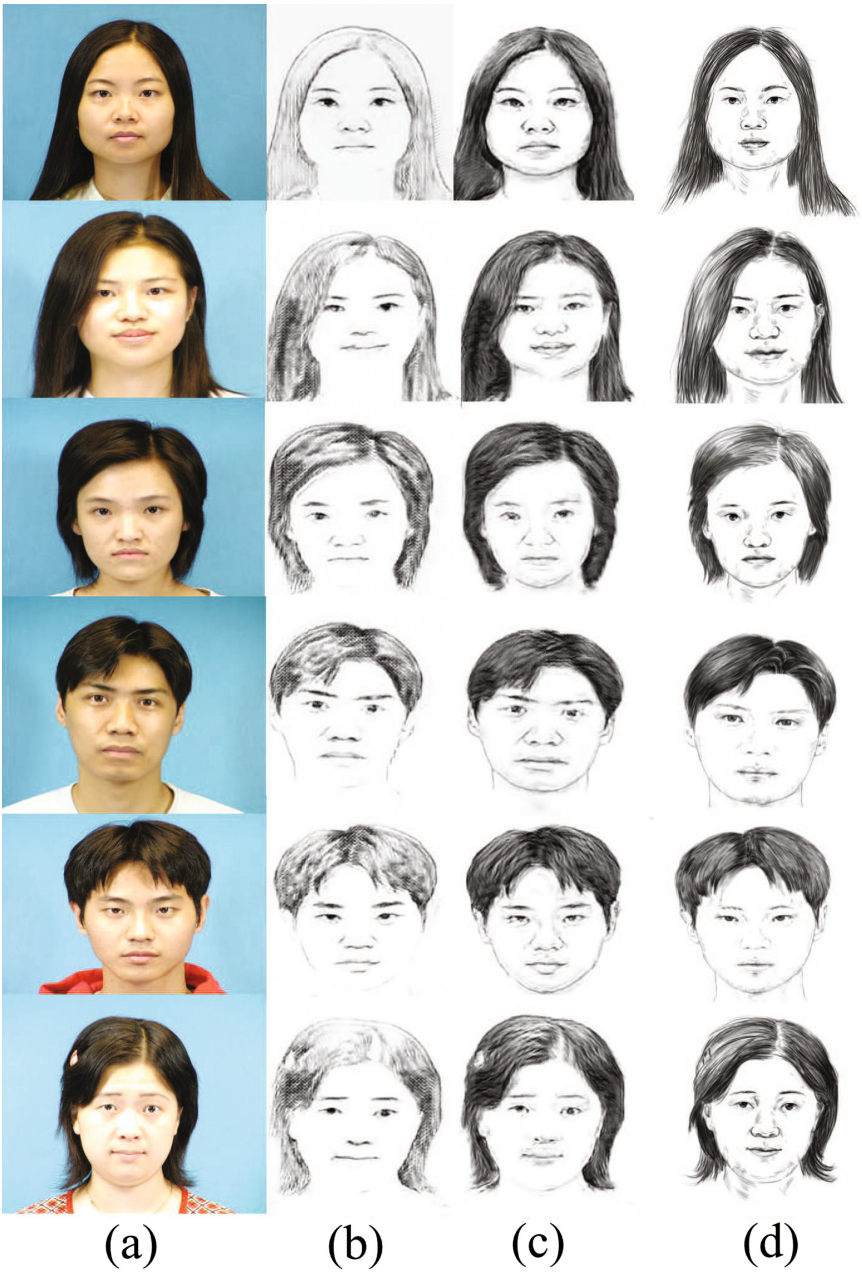

Fig. 5. (a) photos; (b) stage 1 results; (c) stage 2 results(our model); (d) sketches by our artist

5 Conclusion

In this paper, we proposed based on arts drawing steps constrain conditional generation adversarial networks (CGAN). In the first stage, we take the original face photos which concat noise z as input, generating stage 1 low resolution synthesize sketches. In the second stage take the stage 1 results and original face photo as inputs, yielding stage 2 high resolution synthesize sketches, which can express more natural textures and details. Proved by qualitative and quantitative results, our approach dramatically improves the realism of the synthesized face photos and sketches than most previous methods. Also show that two-stage generation is better than the one-stage generation. Future work includes extending this framework to three-stage, four-stage etc.

References

1. Chen, T., Cheng, M.M., Shamir, A., Shamir, A., Hu, S.M.: Sketch2photo: internet image montage. In: ACM SIGGRAPH Asia (2009)
2. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. Computer Science (2014)
3. Di, X., Patel, V.M.: Face synthesis from visual attributes via sketch using conditional VAEs and GANs. arXiv preprint arXiv:1801.00077 (2017)
4. Eitz, M., Richter, R., Hildebrand, K., Boubekeur, T., Alexa, M.: Photosketcher: interactive sketch-based image synthesis. IEEE Comput. Graph. Appl. **31**(6), 56 (2011)
5. Gao, X., Zhong, J., Jie, L., Tian, C.: Face sketch synthesis algorithm based on e-hmm and selective ensemble. IEEE Trans. Circuits Syst. Video Technol. **18**(4), 487–496 (2008)
6. Goodfellow, I.J., et al.: Generative adversarial nets. In: International Conference on Neural Information Processing Systems (2014)
7. Han, Z., Tao, X., Li, H.: Stackgan: text to photo-realistic image synthesis with stacked generative adversarial networks. In: Computer Vision and Pattern Recognition (2016)
8. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Computer Vision and Pattern Recognition (2016)
9. Iwashita, S., Takeda, Y., Onisawa, T.: Expressive facial caricature drawing. In: IEEE International Fuzzy Systems Conference (1999)
10. Johnson, J., Alahi, A., Li, F.F.: Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision (2016)
11. Uhl Jr., R.G., Lobo, N.D.V., Kwon, Y.H.: Recognizing a facial image from a police sketch. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (1996)
12. Jun, Y., Shengjie, S., Fei, G., Dacheng, T., Qingming, H.: Towards realistic face photo-sketch synthesis via composition-aided GANs. In: Computer Vision and Pattern Recognition (2017)
13. Konen, W.: Comparing facial line drawings with gray-level images: a case study on PHANTOMAS. In: von der Malsburg, C., von Seelen, W., Vorbrüggen, J.C., Sendhoff, B. (eds.) ICANN 1996. LNCS, vol. 1112, pp. 727–734. Springer, Heidelberg (1996). https://doi.org/10.1007/3-540-61510-5_123

14. Koshimizu, H., Tominaga, M., Fujiwara, T., Murakami, K.: On kansei facial image processing for computerized facial caricaturing system PICASSO. In: IEEE International Conference on Systems (1999)
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
16. Li, C., Wand, M.: Precomputed real-time texture synthesis with markovian generative adversarial networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 702–716. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_43
17. Lin, Z., Lei, Z., Xuanqin, M., David, Z.: FSIM: a feature similarity index for image quality assessment. IEEE Trans. Image Processing **20**(8), 2378 (2011). A Publication of the IEEE Signal Processing Society
18. Liu, Q., Tang, X., Jin, H., Lu, H., Ma, S.: A nonlinear approach for face sketch synthesis and recognition. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2005)
19. Liu, S., Yang, J., Huang, C., Yang, M.H.: Multi-objective convolutional learning for face labeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3451–3459 (2015)
20. Lucic, M., Kurach, K., Michalski, M., Gelly, S., Bousquet, O.: Are GANs created equal? A large-scale study. In: Advances in Neural Information Processing Systems, pp. 700–709 (2018)
21. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: feature learning by inpainting. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
22. Phillips, P.J., Moon, H., Rauss, P., Rizvi, S.A.: The FERET evaluation methodology for face-recognition algorithms. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 137–143. IEEE (1997)
23. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from simulated and unsupervised images through adversarial training. In: Computer Vision and Pattern Recognition (2016)
24. Tang, X., Wang, X.: Face photo recognition using sketch. In: International Conference on Image Processing (2002)
25. Tang, X., Wang, X.: Face sketch synthesis and recognition. In: Proceedings Ninth IEEE International Conference on Computer Vision, pp. 687–694. IEEE (2003)
26. Wang, L., Sindagi, V.A., Patel, V.M.: High-quality facial photo-sketch synthesis using multi-adversarial networks, pp. 83–90 (2018)
27. Wang, N., Zha, W., Jie, L., Gao, X.: Back projection: an effective postprocessing method for GAN-based face sketch synthesis. Pattern Recogn. Lett. **107**, S0167865517302180 (2017)
28. Wang, X., Tang, X.: Face photo-sketch synthesis and recognition. IEEE Trans. Pattern Anal. Mach. Intell. **31**(11), 1955–1967 (2009)
29. Wang, X., Gupta, A.: Generative image modeling using style and structure adversarial networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 318–335. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_20
30. Yoo, D., Kim, N., Park, S., Paek, A.S., Kweon, I.S.: Pixel-level domain transfer. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 517–532. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_31

31. Zhang, S., Ji, R., Hu, J., Gao, Y., Lin, C.W.: Robust face sketch synthesis via generative adversarial fusion of priors and parametric sigmoid. In: IJCAI, pp. 1163–1169 (2018)
32. Zhao, H., Xia, S., Zhao, J., Zhu, D., Yao, R., Niu, Q.: Pareto-based many-objective convolutional neural networks. In: Meng, X., Li, R., Wang, K., Niu, B., Wang, X., Zhao, G. (eds.) WISA 2018. LNCS, vol. 11242, pp. 3–14. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-02934-0_1
33. Zhou, H., Kuang, Z., Wong, K.Y.K.: Markov weight fields for face sketch synthesis. In: IEEE Conference on Computer Vision and Pattern Recognition (2012)
34. Zhou, Y., Berg, T.L.: Learning temporal transformations from time-lapse videos. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 262–277. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_16
35. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE International Conference on Computer Vision (2017)