GENERATIVE ADVERSARIAL MULTI-TASK LEARNING FOR FACE SKETCH SYNTHESIS AND RECOGNITION

Weiguo Wan[1], *Hyo Jong Lee*[1, 2, *]

[1] Division of Computer Science and Engineering, Chonbuk National University, Korea
[2] Center for Advanced Image and Information Technology, Chonbuk National University, Korea

ABSTRACT

Face sketch synthesis and recognition have wide range of applications in law enforcement. Despite the impressive progresses have been made in faces sketch and recognition, most existing researches regard them as two separate tasks. In this paper, we propose a generative adversarial multi-task learning method in order to deal with face sketch synthesis and recognition simultaneously. Our framework is based on generative adversarial networks (GAN), in which an improved deep network named residual dense U-Net is used as generator to synthesize face sketch image and a multi-task discriminator is designed to not only guide the generator to produce more realistic sketch image, but also extract discriminative face feature. In addition, except the common adversarial loss, the perceptual loss and triplet loss are adopted for the learning of generator and discriminator, respectively. Compared with the state-of-the-art methods, the proposed method obtains better results in terms of face sketch synthesis and recognition.

Index Terms— Face sketch synthesis, face sketch recognition, generative adversarial networks, residual dense U-Net, triplet loss

1. INTRODUCTION

Face sketch synthesis and recognition have attracted significant attention in pattern recognition and computer vision recently, due to its wide range of applications in law enforcement agencies [1]. When a crime happens, if only limited information about the suspect is available due to the low quality of surveillance videos or even no video/image clues, a sketch drawn by the artist according to the description of the witnesses is usually taken as the substitute for suspect identification. When obtaining the sketches, the police can narrow down the suspects by retrieving the law enforcements face datasets or surveillance camera footages with the sketches [2].

However, different from homogeneous face images, there are great modality variations between face sketch images and photo images, which may result in poor performance when applying traditional homogeneous face recognition approaches in face sketch recognition. To make the face photo and sketch images in same domain, some researchers focus on developing face image synthesis algorithms which can transform the face photos to sketches or vice versa. Early face sketch synthesis methods are mainly exemplar-based, which construct the target sketch patch for the test photo patch by weighted averaging of several nearest sketch patches from the training photo-sketch pairs [3-5]. However, the exemplar-based methods are time-consuming, and their results always suffer from blur and deformation. Recently, the deep learning techniques have shown exciting effect in image synthesis and style transfer. They have also been used for face sketch synthesis, such as fully-connected network (FCN) [6] and generative adversarial networks (GAN) [7].

Another strategy to reduce the modality discrepancy in face sketch recognition is to extract modality-invariance feature representation from face photo and sketch images. These methods aim to represent face images based on local feature descriptors such as multiscale local binary patterns (MLBP) [8] and histograms average of oriented gradient (HAOG) [9]. However, the local feature extraction-based methods are sensitive on pose and background variations. Deep learning-based face recognition method can mitigate this problem by learning latent embeddings from vast amounts of face data, such as FaceNet [10] and VGG-Face [11], but it is more challenging to employ deep learning for face sketch recognition due to the limited face photo-sketch data, which may result in over-fitting and local minima when training deep networks.

In this paper, a novel face sketch synthesis and recognition method based on GAN is developed to deal with these two challenging tasks at once. To synthesize high-quality sketch image, we proposed a new generator model, which takes advantage of U-Net architecture [12] and residual dense block (RDB) [13]. Moreover, we designed a multi-task deep network which plays the role of discriminator for generator training and feature extractor for face sketch recognition. Experimental results indicated that the proposed method achieved superior performance both on face sketch synthesis and recognition.

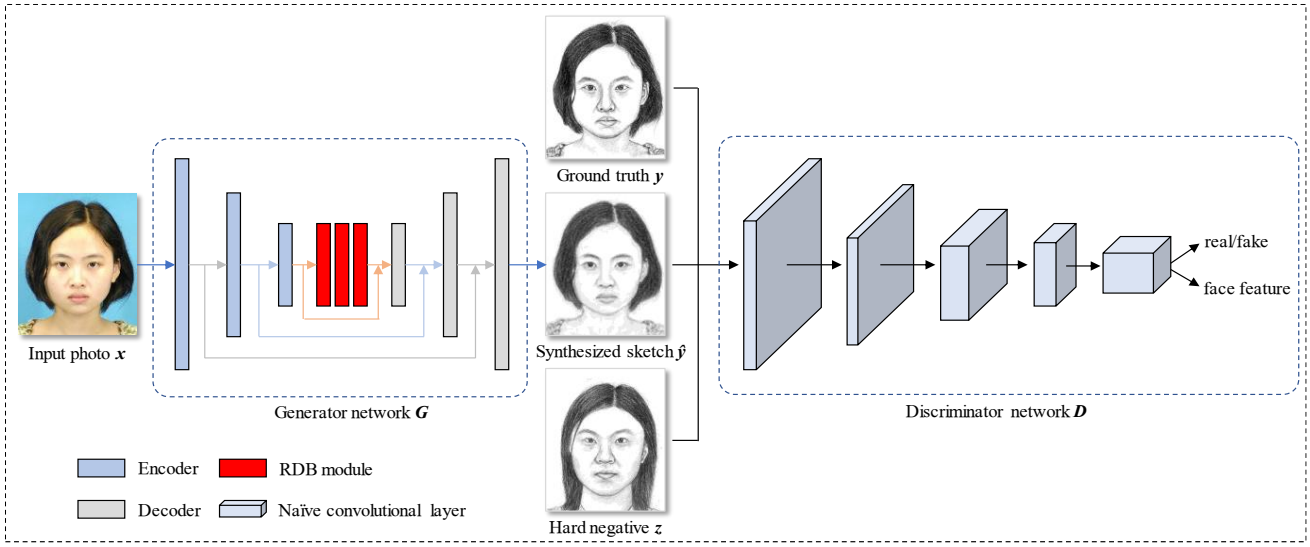

Fig. 1. Overall framework of the proposed face sketch synthesis and recognition method. Feeding a face photo image, a realistic face sketch image is synthesized with the generator network which is composed of residual dense U-Net. The discriminator network learns the discrimination of real/fake sketch image and the feature representation of face sketch image simultaneously.

2. THE PROPOSED FACE SKETCH SYNTHESIS AND RECOGNITION METHOD

In this section, the GAN framework for face sketch synthesis and recognition is proposed. The overall architecture is introduced first, and then the designed generator and discriminator structures will be described in detail. Finally, the loss functions for training the GAN model will be introduced as well.

2.1. Overall architecture

This paper aims at designing a framework that is able to simultaneously synthesize realistic face sketch image with which the domain discrepancy is reduced and extract discriminative face feature for face sketch recognition. The overall architecture (as shown in Fig. 1) mainly consists of two parts: namely a generator (G) and a discriminator (D). The generator is fed with a face photo image and a corresponding sketch image can be obtained. For the discriminator, in order to learn the ability of face feature extraction, three images compose a triplet sample as the input (a generated fake sketch image, a ground-truth sketch image, and a hard-negative sketch image which has small distance with the ground-truth sketch).

2.2. Face sketch synthesis with residual dense U-Net

Though the U-Net is first developed for biomedical image segmentation, its impressive performance in image-to-image synthesis has significantly promoted many other computer vision applications such as shape generation [14], and image deblur [15]. Motivated by its remarkable success, we employ U-Net as the generator of GAN to perform the face sketch synthesis in this paper.

In addition, we modify the U-Net model with the RDB module which can extract abundant local features via dense connected convolutional layers. Different from the original RDB, we conduct the Instance normalization [16] after each convolutional layer in the RDB to improve the quality of the synthesized sketch images. The illustration of the modified RDB module is displayed in Fig. 2. An RDB module consists of dense-connected layers, local feature fusion, and local residual learning three parts, which can be represented as follows:

$$F_{d,i} = \sigma(W_{d,i}[R_{d-1}, F_{d,1}, \ldots, F_{d,i-1}] + b_{d,i}), \quad (1)$$

$$F_{d,LF} = H_d([R_{d-1}, F_{d,1}, \ldots, F_{d,i}, \ldots, F_{d,I}]), \quad (2)$$

$$R_d = R_{d-1} + F_{d,LF}. \quad (3)$$

where $F_{d,i}$ is the output of i-th convolutional layer of d-th RDB module R_d. The σ is the Instance normalization and rectified linear unit (ReLU) activation function. The $W_{d,i}$ and $b_{d,i}$ are the weights and bias of the i-th convolutional layer. The $F_{d,LF}$ and H_d denote the fused local feature maps and the function of 1×1 convolutional layer in d-th RDB.

The input of the generator is a photo image with size of 224×224 and followed by an encoder process composes of three convolutional layers with kernel size 3, zero-padding 1, Instance normalization, and ReLU. The number of filters of these three convolutional layers in the encoder networks are 32, 64, and 128, respectively. The stride of the first layer is 1 and other two layers are 2. After the encoder process, three RDB modules are concatenated and fed into a 1×1 convolutional layer. The output is regarded as the input of the decoder networks, which is the inverse process of the encoder. Corresponding to the encoder networks, three

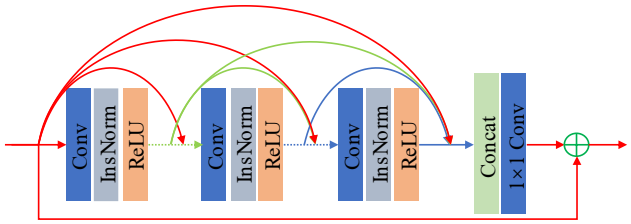

Fig. 2. Architecture of a residual dense block.

convolutional layers are conducted in the decoder networks, and between them two up-sampling operators are adopted. The number of filters of the convolutional layers in the decoder networks are 128, 64, and 32, respectively. Finally, a 3×3 convolutional layer with one filter is used to obtain the target face sketch image.

2.3. Face sketch recognition with discriminator network

The basic structure of the discriminator network consists of five convolutional layers with kernel size 3, padding 1, and strides 2, as shown in Fig. 1. The numbers of filters are 32, 32, 64, 64, and 128, respectively. After each convolutional layer, the batch normalization and ReLU activation are stacked. At the end of the basic discriminator network, a feature map with size of 7×7×128 can be obtained.

Connected to the basic discriminator network, two branches are designed to implement the functions of real/fake sketch discrimination and face feature extraction. The discrimination branch is a convolutional layer with output size of 7×7 and a sigmoid activation layer to predict probability scores between 0 and 1, which is utilized to distinguish the input sketch image is true or fake. The face feature extraction branch is a fully-connected layer with output size of 1024, with which a face sketch image can be represented as a 1024-dimension feature vector. In order to address the lack of training photo-sketch data, we adopt the triplet loss [10] to train the feature extraction branch, which will be introduced in the following section.

2.4. Loss functions

Assume x and y be the training photo-sketch pair, and $\hat{y}$ be the synthesized sketch. The formulas of the loss functions used for training the proposed GAN model will be introduced in this section.

2.4.1. Loss function for generator

Face sketch synthesis can be regarded as a kind of style transfer [17]. Thus, the synthesized sketch should have sketch style and preserve the content of the photo image. The perceptual loss [18] has been improved to be an effective way to deal with the style by extracting the high-level features from deep networks like VGG-19 [19]. The perceptual loss can be expressed as:

$$L_p(y,\hat{y}) = \sum \left\| \Phi_l(y) - \Phi_l(\hat{y}) \right\|_2^2, \tag{4}$$

where Φ_l is the l-th feature map of the VGG-19 network.

In order to preserve the face content, the Charbonnier loss is adopted:

$$L_c(y,\hat{y}) = \sqrt{(y-\hat{y})^2 + \varepsilon^2}, \tag{5}$$

where ε is the Charbonnier penalty factor.

In addition, we use the total variation loss to reduce noise and artifact of the synthesized face sketch image:

$$L_t(\hat{y}) = \sum_{m,n} \left(\left(\hat{y}_{m+1,n} - \hat{y}_{m,n} \right)^2 + \left(\hat{y}_{m,n+1} - \hat{y}_{m,n} \right)^2 \right), \tag{6}$$

where $\hat{y}_{m,n}$ is the pixel value at (m, n) of the synthesized face sketch image.

For the adversarial loss, we adopt the least square loss as in [20]:

$$L_a = \mathbb{E}_{x \sim P_{photo}(x)} [(D(G(x)) - 1)^2], \tag{7}$$

The total loss function for training the generator is calculated by weighted averaging the above loss functions:

$$L_G = \lambda_1 L_p + \lambda_2 L_c + \lambda_3 L_t + \lambda_4 L_a. \tag{8}$$

2.4.2. Loss function for multi-task discriminator

The discriminator network is originally used to distinguish the true sketch image and generated pseudo sketch image. In this paper, it is also trained to extract face feature from face sketch image with the triplet loss. Thus, three inputs are required, include an anchor image $\hat{y}$ (synthesized face sketch), a positive image y (ground true face sketch), a negative image z (real face sketch from different identity). To accelerate the convergence and prevent overfitting, three nearest sketch images are selected as hard negative images for each anchor image, in terms of VGG-Face feature with the ground truth sketch image.

The loss function for the discriminator branch is calculated on generated sketch $G(x)$ and real sketch z:

$$L_D = \frac{1}{2} \mathbb{E}_{z \sim P_{sketch}(z)} [(D(z) - 1)^2] + \frac{1}{2} \mathbb{E}_{x \sim P_{photo}(x)} [D(G(x))^2]. \tag{9}$$

The triplet loss is used to learn discriminative face feature that give the face sketches of the same identity a small distance, and a large distance otherwise:

$$L_T(\hat{y}, y, z) = \max(\mathcal{D}(\mathcal{F}(\hat{y}), \mathcal{F}(y)) - \mathcal{D}(\mathcal{F}(\hat{y}), \mathcal{F}(z)) + \alpha, 0). \tag{10}$$

where $\mathcal{D}$ is the Euclidean distance, $\mathcal{F}$ means the face feature extractor, and α is the margin constant.

3. EXPERIMENTAL RESULTS

3.1. Datasets and implementation details

We evaluate the proposed method on two public face sketch datasets: CUFS [3] and CUFSF [21], which comprise 606 and 1143 well-aligned photo-sketch pairs, respectively. For

Fig. 3. The comparison of synthesized sketches by different methods. First three rows are from CUFS dataset and last two rows are from CUFSF dataset.

convenient comparison, we split the datasets to training set and testing set same as most face sketch synthesis papers [2], [5], [22]. That is, 268 and 250 photo-sketch pairs from CUFS and CUFSF datasets are used for training, and the remaining for testing.

Some implementation details about the proposed method and experiment are introduced here. The λ values in (8) are [1, 100, 0.0001, 1], the margin in (10) is set as 2. The training epochs is 50, and the batch size is 8. The Adam optimizer with learning rate 0.0001 is employed for all the networks training.

3.2. Face sketch synthesis results

To evaluate the performance of our GAN model on face sketch synthesis, four existing methods are compared, namely MWF [4], RSLCR [5], FCN [6], and condition GAN (cGAN) [23]. Fig. 3 shows the synthesized sketches by different methods on CUFS and CUFSF datasets. It can be seen that MWF results lose some content like hair part, RSLCR results suffer from blur, FCN results have many noises. The cGAN method can generate clear sketches, but distortion happened. The proposed method obtains superior results, which are vivid and preserve the face detail well.

3.3. Face sketch recognition results

With the proposed generator and discriminator networks, face photo and sketch images can be represented as 1024-

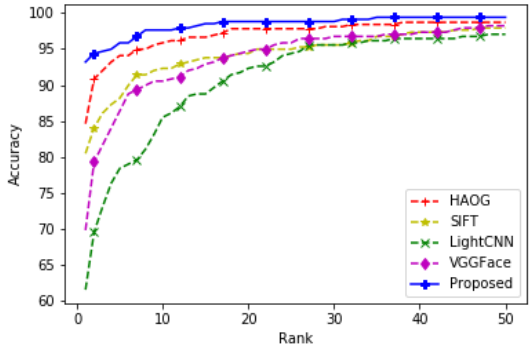

Fig. 4. The comparison of recognition accuracies on CUFS dataset.

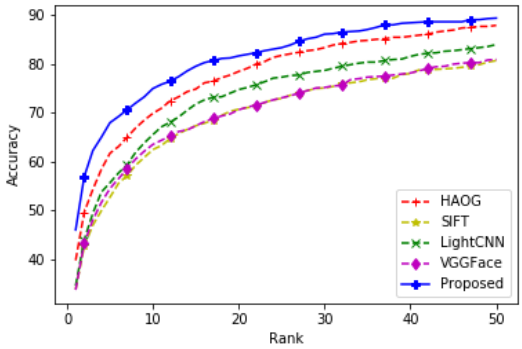

Fig. 5. The comparison of recognition accuracies on CUFSF dataset.

dimension feature vectors, which can be directly used for face similarity measurement. Fig. 4 and Fig. 5 display the performance comparisons of face sketch recognition with extracted face features by our method and other face feature descriptors, include HAOG [9], SIFT [24], LightCNN [25], and VGG-Face [11]. From the recognition performance comparisons, we can find that the proposed method achieves highest accuracies both on CUFS and CUFSF dataset, which indicate the robust and discriminative of the face feature extracted by our method.

4. CONCLUSIONS

In this paper, we presented a novel GAN framework for face sketch synthesis and recognition. A U-Net deep model with residual dense block was designed as generator to synthesize realistic face sketch images. A multi-task discriminator was proposed to train the generator and extract discriminative face feature. Experiments conducted on multiple public datasets demonstrated that the proposed method outperforms the state-of-the-art methods in terms of face sketch synthesis and recognition.

5. ACKNOWLEDGMENTS

This research was supported by Basic Science Research Program through the NRF of Korea funded by the Ministry of Education (GR 2016R1D1A3B03931911).

6. REFERENCES

[1] C. Peng, N. Wang, X. Gao, and J. Li, (2018) "Face recognition from multiple stylistic sketches: Scenarios, datasets, and evaluation," *Pattern Recognition*, vol. 84, pp. 262-272, 2018.

[2] N. Wang, X. Gao, J. Sun, and J. Li, "Anchored neighborhood index for face sketch synthesis," *IEEE Transactions on Circuits System and Video Technology*, vol. 28, no. 9, pp. 2154-2163, 2018.

[3] X. Wang, and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligent*, vol. 31, no. 11, pp. 1955-1967, 2009.

[4] H. Zhou, Z. Kuang, and K. Wong, "Markov weight fields for face sketch synthesis," *in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2012, pp. 1091-1097.

[5] N. Wang, X. Gao, and J. Li, "Random sampling for fast face sketch synthesis," *Pattern Recognition*, vol. 76, pp. 215-227, 2018.

[6] L. Zhang, L. Lin, X. Wu, S. Ding, and L. Zhang, "End-to-end photo-sketch generation via fully convolutional representation learning," *in Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, 2015, pp. 627-634.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, "Generative adversarial nets," *in Advances in neural information processing system*, 2014, pp. 2672-2680.

[8] H. Han, B. Klare, K. Bonnen, and A. Jain, "Matching composite sketches to face photos: a component-based approach," *IEEE Transaction on Information Forensic and Security*, vol. 88, no.1, pp. 191-204, 2013.

[9] H. Galoogahi, and T. Sim, "Inter-modality face sketch recognition," *in Proceedings of International Conference on Multimedia and Expo*, 2012, pp. 224–229.

[10] F. Schroff, D. Kalenichenko, J. Philbin, "Facenet: A unified embedding for face recognition and clustering," *in Proceedings of IEEE conference on Computer Vision and Pattern Recognition*, 2015, pp. 815-823.

[11] O. M. Parkhi, V. Andrea, and Z. Andrew, "Deep face recognition," *in British Machine Vision Conference*, 2015, pp. 1-6.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *in Proceedings of International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234-241.

[13] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," *in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2472-2481.

[14] P. Esser, E. Sutter, B. Ommer, "A variational U-Net for conditional appearance and shape generation," *in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8857-8866.

[15] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, "Deblurgan: Blind motion deblurring using conditional adversarial networks," *in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8183-8192.

[16] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.

[17] L. Gatys, A. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," *in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2414-2423.

[18] J. Johnson, A. Alahi, and F. Li, "Perceptual losses for real-time style transfer and super-resolution," *in European Conference on Computer Vision*, 2016, pp. 694-711.

[19] K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[20] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, S. P. Smolley, "Least squares generative adversarial networks," *in Proceedings of IEEE International Conference on Computer Vision*, 2017, pp. 2813-2821.

[21] W. Zhang, X. Wang and X. Tang, "Coupled information-theoretic encoding for face photo-sketch recognition," *in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 513-520.

[22] M. Zhu, J. Li, N. Wang, X. Gao, "A deep collaborative framework for face photo-sketch synthesis," *IEEE Trans on Neural Networks and Learning Systems*, 2019.

[23] P. Isola, J. Zhu, T. Zhou, and A. Efros, "Image-to-image translation with conditional adversarial networks," *in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5967-5976.

[24] L. Lenc, and K. Pavel, "Automatic face recognition system based on the SIFT features," *Computer and Electrical Engineering*, vol. 46, pp. 256-272, 2015.

[25] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Transaction on Information Forensics and Security*, vol. 13, no. 11, pp. 2884-2896, 2018.