Attributes and Semantic Constrained GAN for Face Sketch-Photo Synthesis

1st Jieying Zheng[1,2,3*], 2nd Haoxian Li[1], 3rd Feng Liu[1]
1. Jiangsu Key Lab on Image Processing & Image Communications, Nanjing University of Posts and Telecommunications, Nanjing, China
2. School of Geographic and Biologic Information,Nanjing University of Posts and Telecommunications
3. Smart Health Big Data Analysis and Location, Nanjing University of Posts and Telecommunications
{zhengjy, 1020010431, liuf}@njupt.edu.cn

2022 14th International Conference on Wireless Communications and Signal Processing (WCSP) | 978-1-6654-5085-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/WCSP55476.2022.10039309

***Abstract*—Face sketch-photo synthesis aims at generating a facial photo conditioned on a given photo. It has attracted wide attention in computer vision and has been widely applied in law enforcement and entertainment. However, precisely synthesizing high-quality face photos is still challenging due to the missing color information in face sketches. To alleviate the problem, we propose an attribute and Semantic constrained Generative Adversarial Networks (ASGAN), which introduces face attribute and semantic constraints to improve the quality and accuracy of synthesized photos. Specifically, We first annotate the key attributes in the dataset and form a 14-dimensional attribute vector for each face. Then, the attribute features and semantic features obtained by face parsing are fused. Finally, the generator takes the fused constrain features and sketches as input and constructs two-stream encoders to synthesize high-quality photos. Extensive experimental results demonstrate that our method can significantly outperform state-of-the-art methods. It can synthesize higher-quality face photos while maintaining the identity, attributes, and structure. Meanwhile, it can alleviate the problem of background and skin color synthesis errors.**

***Index Terms*—face sketch-photo synthesis, attribute, generative adversarial networks, face parsing**

I. Introduction

Face sketch-photo synthesis aims at generating face photos from the corresponding input sketches while ensuring face consistency. It has been widely used in criminal law enforcement and digital entertainment [1]. In past years, lots of works have made a great effort to this task, which can be divided into two categories: exemplar-based [2] and deep learning-based methods [3], [4], [6], [7]. The exemplar-based methods synthesize the target images by dividing images into patches, neighbors search, and weight computation. However, such methods destroy the structure of face and are time-consuming. Deep learning-based methods alleviate these problems. Thanks to the introduction of the Generative Adversarial Networks (GAN) [5], the development of face sketch-photo synthesis has reached a new level. The related image-to-image translation methods [4], [6], [7] also achieve good results on the face sketch-photo synthesis. The main problems of the GAN-based methods are the details of synthesized photos and facial structure maintenance are not satisfactory. Many works try to improve the performance of GAN using multi-scale discriminators [8], stacking strategy [9], collaborate learning [10], global and local synthesis [11].

Fig. 1. Examples from the CUFS database and sketch parsing. For each example, from left to right are sketch, sketch parsing and photo, sequentially.

Recently, many works [12]–[14] try to utilize facial composition information to guide the generation. Yu et al. [12] use face parsing masks and introduce a compositional reconstruction loss to improve the weights of important facial components. Li et al. [13] propose a geometric consistency loss to further minimize the locally structural divergence between the synthesized faces and inputs. In addition, the SPatially Adaptive DEnormalization (SPADE) [15] module always be introduced to the generator to enhance the semantic information. In our work, face parsing is also used to constrain the semantic structure of synthesized photos.

Generally, sketch-photo synthesis and photo-sketch synthesis are considered to be the same task [6], [12], [16]. However, compared to photos, face sketches lack much information, which makes sketch-photo synthesis be a more difficult task. The color information is important for the identification, such as skin tones. In previous works, the missing information mainly relied on the training dataset. Our motivation is to utilize the given attributes to further constrain the generator to synthesize photos precisely.

* Jieying Zheng is the corresponding author. This work is supported in part by the National Natural Science Foundation of China under Grant (No.62177029), in part by the Natural Science Foundation of Nanjing University of Posts and Telecommunications (Grant No. NY221104) and in part by the Startup Foundation for Introducing Talent of Nanjing University of Posts and Communications under Grant No. NY221031.

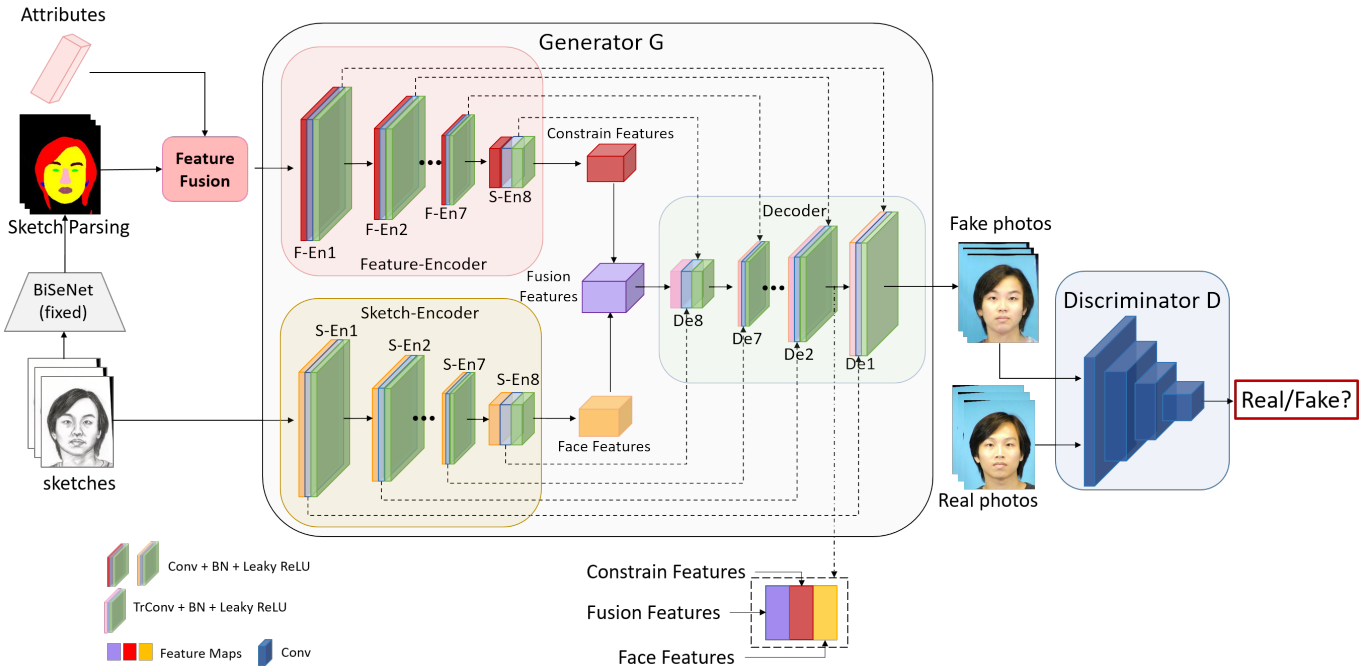

Fig. 2. The framework of our ASGAN method.

To generate face photos with correct attributes and semantic structure, we constrain the generator with attribute vectors and semantic masks. To this end, we propose the Attributes and Semantic Constrained Generative Adversarial Networks, named ASGAN. We annotate the face attributes in the dataset and use face parsing to obtain semantic masks (as shown in Fig. 1). Then, combining sketches, attributes, and face parsing as input to guide the high-quality generation of face photos. The main contributions of this paper can be summarized as follows:

- The 14-dimensional important attributes of the faces in the CUFS database [17] are annotated, such as skin color, hair color, bread, etc.
- Face attribute information is fused with face semantic information, and the fused features are embedded into the generative adversarial networks to guide the synthesis of more accurate face photos.
- Extensive comparative experiments confirm that our method can outperform state-of-the-art methods.

II. Method

Our approach is aimed at sketch-photo synthesis and mainly addresses the problem of attribute errors in the synthesized photos, which is caused by missing information in input sketches. Given a paired sketch-photo face dataset $\{x_i, y_i\}_{i=1}^{N}$, where $x_i \in X$, $y_i \in Y$ represent sketches and photos, and N is the number of image pairs. Given a face sketch x, its semantic p, and a face attribute vector a, our goal is to generate a corresponding face photo y with the same identity and natural appearance for recognition. To achieve this, we propose a GAN-based framework with attribute and semantic constraints (ASGAN). The problem can be expressed mathematically as: $G(x, a, p) \rightarrow \hat{y}$. The framework of our proposed ASGAN method is shown in Fig. 2. As shown in Fig. 2, our framework consists of two networks: a generator G and a discriminator D. While the generator G contains two encoders and one decoder. Encoder $Feature-Encoder$, as part of the generator G, takes the attribute a and face parsing semantic p as input, fuses the two features, and finally embeds the fused features into the generator.

A. Face Attributes

Compared to photos, sketches lack a lot of information, especially color information. Therefore, for face sketch-photo synthesis, some information is uncertain, which leads to the problem of incorrect in synthesizing some attributes such as skin color. However, it is important for identification. To solve this problem and improve the synthesis performance of the model, our key idea is to supplement import missing information for the input with annotated attributes. The commonly used CUFS dataset [17] has not been annotated in previous works. Therefore, we first annotated the color attributes of face images in the CUFS dataset.

According to the face photos in the CUFS dataset, We annotated the background, skin, hair, eyes and beard colors of all faces in the CUFS dataset. And each face forms a corresponding 14-dimension vector. Where, background, skin, hair, eyes, and beard occupy 3, 3, 4, 2, and 2 dimensions, respectively. Therefore, an attribute vector can be $a = [0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 1, 0, 1, 0]$ as example.

B. Face Parsing

The semantic structure of the faces are obtained by face parsing methods. We utilize the pre-trained BiSeNet [18] trained on the CeleAMask-HQ dataset [19] to obtain the face parsing masks of input photos. The output of BiseNet contains pixel-wise labels related to 19 components, including background, skin, left eye, right eye, left eyebrow, etc. We

merged similar components, e.g. left eye and right eye, and finally form 8 key components, including background, facial skin, two eyebrows, two eyes, nose, inner mouth, upper and lower lips, and hair. Semantic parsing masks of all faces in the CUFS database are obtained before training to save time. The semantic parsing can be defined as $P = \{P^1, P^2, \ldots, P^8\}$.

C. Feature Fusion

After obtaining attributes and semantic features, the features of different modalities must be fused and a feature fusion module is introduced. For attribute features, we first expand the attribute vectors to the size of sketch parsing. Then directly concatenate the extended attribute vector with the sketch parsing in the channel direction. Finally, the concatenated features are passed through to a Convolution-BatchNorm-LeakyReLU module and a residual module as the fused constrain features. The feature-encoder mainly encodes the constraint features of attributes and semantics. The sketch-encoder encodes facial features in face sketches.

D. Network Architecture

Inspired by the U-Net [20] framework, the architecture of Generator G is an encoder-decoder structure. But the difference is that the encoding of G is a two-stream structure. As shown in Fig. 2, it has two encoders: Feature-Encoder and Sketch-Encoder. They both contain 8 encoding layers (F-En1$\sim$8 and S-EN1$\sim$8), each encoding layer consists of a convolutional layer, a Batch-Normalization layer and a Leaky ReLU layer (Conv-BN-LeakyRelu). Correspondingly, the decoder contains 8 decoder layers (De8$\sim$1) and each encoding layer consists of a transposed convolutional layer, a Batch-Normalization layer and a leaky ReLU layer (TrConv-BN-LeakyRelu).

The discriminator D consists of four convolutional layers. First three layer are followed by a Batch-Normalization layer and a leaky ReLU layer. The last layer predicts a matrix of real/fake labels of the input photo.

In the training stage, D tries to correctly distinguish fake photos from real sketches; while G tries to fool D by generating photos as real as possible. By alternately updating the parameters of G and D until a Nash equilibrium is reached, the final model is obtained.

E. Objective Function

Adversarial Loss: Since our model adopts the framework of Generative Adversarial Networks, the adversarial loss is defined as:

$$L_{adv} = \log D(X, E, Y) + \log\left(1 - D(X, E, \hat{Y})\right) \quad (1)$$

The patchGAN [4] is used in the discriminator D.

***Compositional Loss*:** Conditional GAN [21] based image synthesis tasks are always trained with pixel-wise reconstruction loss and the normalized ℓ_1 distance encourages less blurring. Therefore, We compute the global reconstruction loss L_{global} using the ℓ_1 distance between the synthesized photos $\hat{Y}$ and the target photos Y.

In the global reconstruction loss, components with more pixels contribute more. But it doesn't match the reality. For example, the background contains a large number of pixels but has little impact on performance. In the generator, we utilize face parsing to constrain face components. Following CAGAN [12], we employ a local reconstruction loss L_{local} to balance the effects of eight components with different pixel counts. In practice, we weighted the global reconstruction loss and compositional reconstruction loss as the final component loss L_{cmp}:

$$\begin{aligned} L_{cmp} &= \alpha L_{global} + (1-\alpha) L_{local} \\ &= \alpha \frac{\|Y - \hat{Y}\|_1}{hw} + (1-\alpha) \sum_{c=1}^{8} \frac{\|Y \odot P^c - \hat{Y} \odot P^c\|_1}{P^c \otimes 1} \end{aligned} \quad (2)$$

Where, h, w are the height and width of the photos, $\otimes$ denotes the convolution operation and $\odot$ is the pixel-wise product operation. P^c is the c-th component of the sketch paring P, and the balance factor α is set to 0.7.

Perceptual Loss: To ensure the synthesized photos are more realistic and natural, we introduce perceptual loss to constrain the synthesized photos. The perceptual loss is the feature difference extracted from different layers of the VGG network between the synthesized photos and the target photos. It can be formulated as:

$$L_{per} = \sum_{l \in S} \left\| \varphi^l(Y) - \varphi^l(\hat{Y}) \right\|_2 \quad (3)$$

Where $\varphi^l(\cdot)$ represents the feature output of the l layer in the VGG-16 network [22]. We select Conv-1, Conv3-4, and Conv5-1 as the feature set S, which includes low-level, mid-level and high-level features.

Full Objective: Finally, all the losses mentioned above are combined to train our ASGAN:

$$(G^*, E^*, D^*) = arg \min_{G,E} \max_{D} L_{adv} + \lambda L_{cmp} + \gamma L_{per} \quad (4)$$

Where, λ and γ are the weight parameter to balance the importance of different losses, set to 10 and 5, respectively. In training, the full objective function is optimized by solving the max-minimization problem and updating the parameters of generator and discriminator alternately .

III. Experiments

In this section, we evaluate the performance of our proposed method experimentally.

A. Settings

The widely used CUFS database [17] is used to evaluate the performance of different methods. The CUFS database consists of 606 well-aligned face photo-sketch pairs, including the CUHK student dataset (188 pairs), the AR dataset (123 pairs), and the XM2VTS dataset (295 pairs). The samples of the database are shown in Fig. 1. Although the photos are different, we train them together in our method, with 268 face photo-sketch pairs used for training and the rest 338 pairs for testing. In training, the Adam optimizer with $\beta_1 = 0.5$

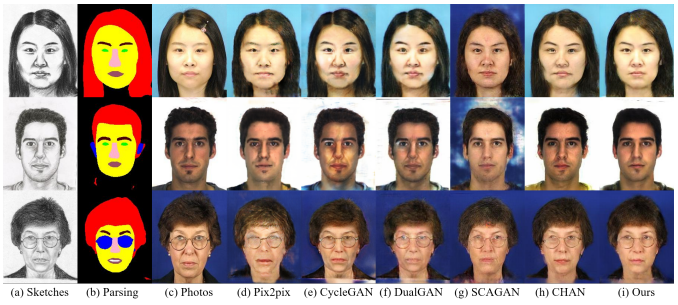

Fig. 3. Examples of synthesized face photos on the CUFS dataset.

and $\beta_2 = 0.999$ is used, the learning rate is set to 0.00002 and batch size to 1. Our model is trained for 700 epochs on a single NVIDIA 2080Ti GPU. Five state-of-the-art methods are selected for performance comparison: Pix2pix [4], CycleGAN [6], DualGAN [7], SCAGAN [12] and CHAN [16] methods. We obtain the synthesized results or source codes directly from the authors to avoid biases. For quantitative comparison, we choose Feature Similarity (FSIM) [23], Learned Perceptual Image Patch Similarity (LIPIPS) [24], and the Null-space Linear Discriminant Analysis (NLDA) [25] based face recognition rates as the evaluation metrics. The LPIPS values are calculated with three networks: Squeeze, Alex and VGG.

B. Qualitative Results

The face photo synthesized results of different methods are shown in Fig. 3. We can see that our method achieves the best visual performance. Our resulted photos contain richer details and identity information. Moreover, our method can effectively alleviate the problem of background and skin color synthesis mistakes caused by the loss of color information. It illustrates the effectiveness of attribute constraints in the network. Meanwhile, the embedding and fusion of semantic information also help to synthesize faces with more accurate structures.

C. Quantitative Results

The quantitative comparison results are shown in TABLE I. As can be seen from the table, our method achieves the best performance on all metrics with significant improvements. It shows that our method is more capable of synthesizing the highest quality face photos among all the compared methods.

TABLE I. COMPARISON WITH STATE-OF-THE-ART METHODS ON THE CUFS DATASET. THE BEST AND SECOND BEST OF EACH METRICS WILL BE HIGHLIGHTED IN BOLDFACE AND UNDERLINE FORMAT, RESPECTIVELY. ↓ INDICATES THE LOWER IS BETTER, AND ↑ HIGHER IS BETTER.

Methods	**FSIM** ↑	**LPIPS** ↓ (alex)	**LPIPS** ↓ (squeeze)	**LPIPS** ↓ (vgg)	**NLDA**↑ (%)
Pix2pix [4]	0.7678	0.1981	0.1769	0.3256	<u>92.75</u>
CycleGAN [6]	0.7682	0.1876	0.1685	0.3279	86.83
DualGAN [7]	0.7716	0.2112	0.1846	0.3210	92.38
SCAGAN [12]	0.7764	0.1790	<u>0.1539</u>	<u>0.3011</u>	89.38
CHAN [16]	<u>0.7781</u>	*0.1780*	0.1553	0.3044	87.50
ASFS (Ours)	**0.7845**	**0.1635**	**0.1433**	**0.2849**	**96.77**

Meanwhile, Fig. 4 shows the face recognition accuracy (NLDA) with the variation of the reduced number of dimensions.

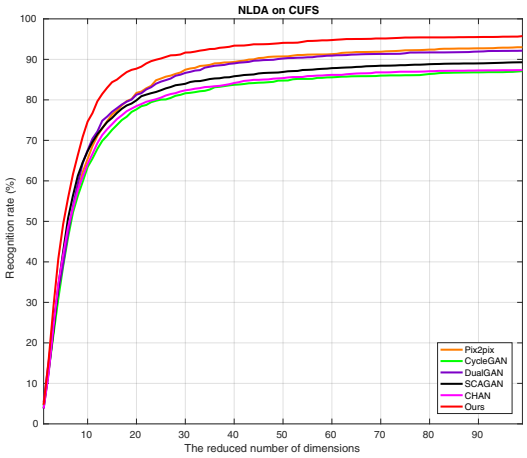

Fig. 4. NLDA face photo recognition accuracy with the variation of the reduced number of dimensions.

It can be seen from both Fig. 4 and TABLE I, our method can obtain the highest face recognition rate of 96.77% among all methods. It also illustrates that the face photos synthesized by our method are more accurate, and the constraints of attributes and semantics are meaningful for synthesizing more accurate faces.

D. Ablation Study

To prove the effectiveness of our method, we conduct extensive ablation experiments to analyze the contribution of different constraints in the proposed model and different losses in the objective function to model performance.

To demonstrate the effectiveness of attribute and semantic constraints, we separate attributes and semantic constrain modules from the full model and conduct comparison experiments. The results of ablation experiments are shown in TABLE II. Obviously, both semantic and attribute constraints play important roles in improving model performance. However, it can be seen that attributes contribute more to model performance than semantics, especially in preserving identity information and improving the face recognition rate.

TABLE II. ABLATION EXPERIMENTS ON THE CUFS DATASET.

Models	**FSIM** ↑	**LPIPS** ↓ (alex)	**LPIPS** ↓ (squeeze)	**LPIPS** ↓ (vgg)	**NLDA**↑ (%)
base model	0.7708	0.1890	0.1630	0.3140	87.54
base model+semantic	0.7736	0.1826	0.1588	0.3066	90.02
base model+attributes	<u>0.7828</u>	<u>0.1698</u>	<u>0.1467</u>	<u>0.2886</u>	<u>93.25</u>
full model	**0.7845**	**0.1635**	**0.1433**	**0.2849**	**96.77**

In addition, we analyze the contribution of different losses in the objective function to the model performance. TABLE. III shows the quantization results using different losses in training. As can be seen from the table, all losses: L_{adv}, L_{cmp} and L_{per} play an important role in training. Compositional loss L_{cmp} focuses on maintaining identity information, which is more conducive to improving the face recognition rate. The perceptual loss L_{per} focuses more on improving the subjective quality of the synthesized photos, which is beneficial to reduce the values of LPIPS. When the full loss is used, the synthesis quality and face recognition rate can achieve a better balance and obtain better metric values.

(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Fig. 5: Ablation experiments of our ASGAN on the CUFS dataset: (a) sketch, (b) sketch parsing, (c) target photo, (e) base model, (h) base model+semantic, (i) base model+attributes, (g) L_{adv}, (h) $L_{adv} + L_{cmp}$, (i) $L_{adv} + L_{per}$, (j) full model with full loss.

TABLE III. ABLATION EXPERIMENTS ON THE CUFS DATASET.

Losses	**FSIM** ↑	**LPIPS** ↓ (alex)	**LPIPS** ↓ (squeeze)	**LPIPS** ↓ (vgg)	**NLDA**↑ (%)
L_{adv}	0.7524	0.2387	0.2239	0.4069	86.44
$L_{adv} + L_{cmp}$	0.7809	0.1710	0.1486	0.2991	**96.97**
$L_{adv} + L_{per}$	0.7831	**0.1619**	**0.1417**	0.2870	93.01
full loss	**0.7845**	0.1635	0.1433	**0.2849**	96.77

The synthesized results corresponding to different models and losses are shown in Fig. 5. The last is result of full model with full loss. The figure illustrates the effectiveness of each module and each loss in our model.

IV. CONCLUSION

In this paper, we proposed a face sketch-photo synthesis method with attributes and semantics constrained generative adversarial networks (ASGAN). Specifically, we annotated the face attributes of the CUFS dataset and formed 14-dimensional attribute feature vectors. To generate face photos with correct attributes and structures, attribute features and semantic features of face parsing are fused. Meanwhile, the fusion features are encoded and embedded in the generator to guide the synthesis of face photos. Extensive experimental results demonstrate that our method can outperform state-of-the-art methods and synthesize face photos with correct attributes. In future work, we will explore the multi-modal face sketch-photo high-quality synthesis with the constraints of face attributes.

REFERENCES

[1] X. Tang and X. Wang, "Face sketch recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 50–57, Jan 2004.

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

[3] L. Zhang, L. Lin, and X. W. et al., "End-to-end photo-sketch generation via fully convolutional representation learning," in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, ser. ICMR '15. New York, NY, USA: ACM, 2015, pp. 627–634.

[4] P. Isola, J. Zhu, and T. Z. et al., "Image-to-image translation with conditional adversarial networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, Hawaii, USA, July 2017, pp. 5967–5976.

[5] I. J. Goodfellow, J. P. Abadie, and M. M. et al., "Generative adversarial nets," in *NIPS*, 2014.

[6] J. Zhu, T. Park, and P. I. et al., "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 2242–2251.

[7] Z. Yi, H. Zhang, and P. T. et al., "Dualgan: Unsupervised dual learning for image-to-image translation," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 2868–2876.

[8] L. Wang, V. Sindagi, and V. Patel, "High-quality facial photo-sketch synthesis using multi-adversarial networks," in *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, May 2018, pp. 83–90.

[9] H. Zhang, T. Xu, and H. L. et al., "Stackgan++: Realistic image synthesis with stacked generative adversarial networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1947–1962, 2019.

[10] M. Zhu, J. Li, N. Wang, and X. Gao, "A deep collaborative framework for face photosketch synthesis," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 10, pp. 3096–3108, 2019.

[11] M. Zhang, R. Wang, X. Gao, J. Li, and D. Tao, "Dual-transfer face sketchphoto synthesis," *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 642–657, 2019.

[12] J. Yu, X. Xu, and F. G. et al., "Toward realistic face photosketch synthesis via composition-aided gans," *IEEE Transactions on Cybernetics*, vol. 51, no. 9, pp. 4350–4362, 2021.

[13] X. Li, F. Gao, and F. Huang, "High-quality face sketch synthesis via geometric normalization and regularization," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 2021, pp. 1–6.

[14] D. Zhang, L. Lin, and T. C. et al., "Content-adaptive sketch portrait generation by decompositional representation learning," *IEEE Transactions on Image Processing*, vol. 26, no. 1, pp. 328–339, 2017.

[15] T. Park, M. Liu, and T. W. et al., "Semantic image synthesis with spatially-adaptive normalization," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2332–2341.

[16] F. Gao, X. Xu, and J. Y. et al., "Complementary, heterogeneous and adversarial networks for image-to-image translation," *IEEE Transactions on Image Processing*, vol. 30, pp. 3487–3498, 2021.

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

[18] C. Yu, J. Wang, and C. P. et al., "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[19] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

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

[21] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *Computer Science*, pp. 2672–2680, 2014.

[22] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.

[23] Zhang, Lin, and Z. et al., "Fsim: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[24] Zhang, Richard, and I. et al., "The unreasonable effectiveness of deep features as a perceptual metric," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

[25] L. F. Chen, H. Y. M. Liao, and M. T. K. et al., "A new lda-based face recognition system which can solve the small sample size problem," *Pattern Recognition*, vol. 33, no. 10, pp. 1713–1726, 2000.