2022 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-9620-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICIP46576.2022.9897523

FACE PHOTO SYNTHESIS VIA INTERMEDIATE SEMANTIC ENHANCEMENT GENERATIVE ADVERSARIAL NETWORK

Haoxian Li[1] *Jieying Zheng*[1] *Feng Liu*[1,⋆]

[1]Jiangsu Key Lab of Image Processing & Image Communications,
Nanjing University of Posts and Telecommunications

ABSTRACT

Face sketch-photo synthesis is an important task in computer vision now. Recently, researchers have introduced face parsing to further improve the quality of synthesized face images. However, the semantic difference between face sketch parsing and photo parsing is usually ignored, leading to deformations and aliasing on synthesized face images. To solve these problems, we propose an **intermediate face parsing** to enhance the semantic information of the input face parsing. According to this intermediate face parsing, we propose an **Intermediate Semantic Enhancement Generative Adversarial Network (ISEGAN)** to generate high-quality realistic face photos. Furthermore, a **Parsing Matching Loss (PM Loss)** is proposed to encourage the intermediate face parsing to be more semantically accurate. Extensive comparison experiments demonstrate that our ISEGAN significantly outperforms the state-of-the-art methods.

***Index Terms*—** deep learning, face sketch-photo synthesis, generative adversarial network, face parsing

1. INTRODUCTION

Face sketch-photo synthesis, commonly understood as synthesizing a face photo from its corresponding sketch, has made great progress in the past few years. It is widely used in digital entertainment and law enforcement [1]. Recently, inspired by the great achievements of Generative Adversarial Network (GAN) [2] in image-to-image translation tasks [3–5], researchers proposed numerous interesting GAN-based methods which drive the development of face sketch-photo synthesis. Wang et al. [6] proposed a multi-adversarial network to iteratively generate high-resolution face images from low-resolution face images using multi-scale discriminators. Ji et al. [7] used multi-domain adversarial learning to learn the mapping between photos and sketches. Zhu et al. [8] used the knowledge from a high-performance large network to strengthen the performance of a light-weight network. Yu et al. [9] decomposed the face parsing [10] into multiple compositions and encoded them into semantic features for face sketch synthesis. Moreover, the stacking strategy [11] was used in [9] as a refinement trick to further improve the performance. Li et al. [12] introduced the SPADE [13] module and adopted spatially-adaptive normalization on latent features according to face parsing layouts. Gao et al. [14] incorporated two complementary heterogeneous generators to generate the coarse face and facial details respectively.

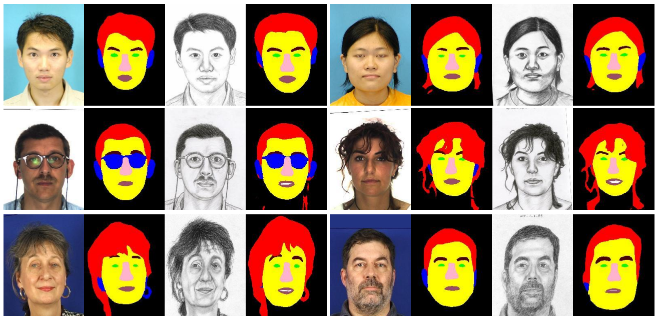

Fig. 1. Examples from the CUFS dataset and their corresponding parsing. For each example, from left to right are photo, photo parsing, sketch and sketch parsing, sequentially.

Using face parsing prior is a popular and effective trick in face sketch-photo synthesis now. Face parsing facilitates visually comfortable face images with clear facial components and realistic texture in [9, 12, 15]. However, after conducting multiple rigorous experiments, we observe that the performance of these methods in face photo synthesis is not so good as that in face sketch synthesis. Specifically, obvious deformations and aliasing defects still exist over the mouth and eyes regions of synthesized photos, or even heavy noise and artifacts. Inspired by our observation and follow-up contrast experiments, we consider that there is a semantic mismatching between face sketch parsing and photo parsing. In other words, it is this mismatching that causes the above defects.

To solve these problems, we propose an intermediate face parsing to enhance semantic information of the input face parsing. Especially, we propose an Intermediate Semantic Enhancement Generative Adversarial Network (ISEGAN) which exploits this intermediate face parsing to synthesize high-quality realistic face photos. Additionally, a parsing

⋆ Feng Liu is the corresponding author. This work is supported by the National Natural Science Foundation of China(No.62177029).

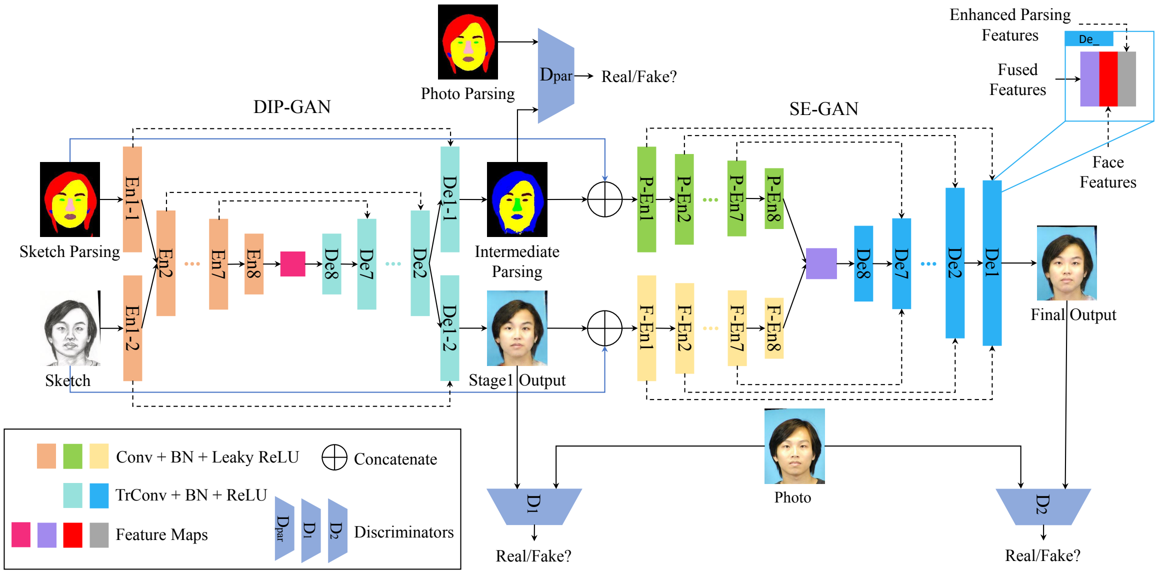

Fig. 2. The framework of our ISEGAN.

matching loss is proposed to improve the synthesized intermediate face parsing by balancing the contributions of the global structure-level semantic information and the channel-wise class-level semantic information. The contributions of this paper are summarized as follows:

- We propose an intermediate face parsing to solve the problems caused by the semantic mismatching between face sketch parsing and photo parsing.
- An Intermediate Semantic Enhancement Generative Adversarial Network (ISEGAN) is proposed for high-quality realistic face photo synthesis.
- An parsing matching loss is proposed to improve the intermediate face parsing.
- We conduct extensive comparison experiments and our ISEGAN outperforms the state-of-the-art methods.

2. METHOD

2.1. Overview

The overview of our ISEGAN is shown in Fig. 2. First, we design a Dual-discriminator Intermediate Parsing Generative Adversarial Network (DIP-GAN) in stage-I to learn the mapping between the sketch domain and the photo domain. The DIP-GAN aims at generating an intermediate face parsing P_{int} and a preliminary photo $\hat{Y}^{(1)}$. Then, we concatenate outputs and inputs of DIP-GAN as shown in Fig. 2. The concatenating operation on sketch parsing P_s and intermediate parsing P_{int} can be regarded as a semantic enhancement, obtaining an enhanced parsing. Finally, we propose a Semantics Embedding Generative Adversarial Network (SE-GAN) in stage-II. The SE-GAN encodes the enhanced parsing into semantic features and fuses them with latent face features, generating high-quality realistic face photo $\hat{Y}^{(2)}$. More details of our method are as follows.

Face parsing. The face sketch parsing P_s and photo parsing P_p are obtained from [10] before training. All face sketches, photos and their corresponding face parsing have the same size $h \times w$ and some examples are presented in Fig. 1. Each face parsing involved in our method contains eight key components, including background, facial skin, two eyebrows, two eyes, nose, inner mouth, upper and lower lips, and hair. All face parsing in our method can be uniformly defined as $P = \{p^{(1)}, p^{(2)}, ..., p^{(8)}\}$.

2.2. Dual-discriminator Intermediate Parsing Generative Adversarial Network (DIP-GAN)

Inspired by the dual-discriminator GAN in [16] and the generator in [12], we design DIP-GAN as the stage-I network to learn the mapping between the sketch domain and the photo domain. The DIP-GAN includes a generator G_1, a parsing discriminator D_{par} and a photo discriminator D_1.

Generator: The G_1 is a "U-Net" shape generator [3], consisting of 8 encoding blocks and 8 decoding blocks. The G_1 encodes the sketch X and its corresponding sketch parsing P_s respectively, and fuses these two kinds of features as the latent representation. In order to generate the intermediate parsing P_{int} and preliminary photo $\hat{Y}^{(1)}$ from the latent representation, we design two branches in the last decoding block, marked as De1-1 and De1-2. The output layer of De1-1 is a Softmax layer for generating face parsing and that of De1-2 is a Tanh layer for generating face photos. Besides, we concatenate the outputs of the i-th encoding block En(i) with those of the (i+1)-th decoding block De(i+1).

Discriminators: D_{par} makes P_{int} close to the photo parsing P_p and D_1 makes $\hat{Y}^{(1)}$ close to the target Y. Both D_{par} and D_1 follow the 70×70 PatchGAN discriminator in [3].

2.3. Semantics Embedding Generative Adversarial Network (SE-GAN)

The SE-GAN contains a generator G_2 and a discriminator D_2.

Generator: G_2 is also a "U-Net" shape generator [3], formed by a parsing encoder, a face encoder, and a decoder. G_2 uses the parsing encoder and the face encoder to encode its two inputs, respectively. To strengthen the structure and texture information of synthesized face photos, the output features of the i-th encoding block P-En(i) in parsing encoder and F-En(i) in face encoder will be concatenated with the outputs of the (i+1)-th decoding block De(i+1). The output layer of the last decoding block De1 is a Tanh layer.

Discriminator: D_2 is a 16×16 PatchGAN discriminator, aiming to make the final output $\hat{Y}^{(2)}$ close to the target Y.

2.4. Objective Function

Adversarial loss in stage-I. According to the dual-discriminator adversarial loss defined in [16], our adversarial loss in stage-I can be divided into $L_{adv1-face}$ and $L_{adv1-par}$. To balance the focus of DIP-GAN on face and parsing, we use a weighted average of these two items. And the loss is given as

$$\begin{aligned} L_{adv1} &= \alpha L_{adv1-face} + (1-\alpha) L_{adv1-par} \\ &= \alpha E_{X,P_s,Y} \left[\log D_1 \left(X, P_s, Y\right)\right] \\ &+ \alpha E_{X,P_s} \left[\log \left(1 - D_1 \left(X, P_s, \hat{Y}^{(1)}\right)\right)\right] \\ &+ (1-\alpha) E_{X,P_s,P_p} \left[\log D_{par} \left(P_p\right)\right] \\ &+ (1-\alpha) E_{X,P_s} \left[\log \left(1 - D_{par} \left(P_{int}\right)\right)\right] \end{aligned} \tag{1}$$

where the weight factor α is set to 0.3.

Adversarial loss in stage-II. Following the setting of [3], the adversarial loss in stage-II can be formulated as

$$\begin{aligned} L_{adv2} &= E_{X,P_s,Y} \left[\log D_2 \left(X, P_s, Y\right)\right] \\ &+ E_{X,P_s} \left[\log \left(1 - D_2 \left(X, P_s, \hat{Y}^{(2)}\right)\right)\right] \end{aligned} \tag{2}$$

Parsing matching loss. To balance the contributions of the global structure-level semantic information and the channel-wise class-level semantic information during the generation of the intermediate parsing, we propose the parsing matching loss (PM loss). The PM loss is composed of a global matching item and a channel matching item. In practice, we use a weighted average of these two items, and the PM loss can be expressed as

$$L_{PM} = \beta \left\| P_p - P_{int} \right\|_2 + (1-\beta) \sum_{c=1}^{8} \left\| p_p^{(c)} - p_{int}^{(c)} \right\|_2 \tag{3}$$

where the weight factor β is set to 0.3.

Compositional loss. To generate more accurate facial components, we substitute our intermediate face parsing for the sketch parsing of the original loss in [9]. The compositional loss in our method can be expressed as

$$L_{cmp} = \delta \frac{\left\| Y - \widehat{Y} \right\|_1}{hw} + (1-\delta) \sum_{c=1}^{8} \frac{\left\| Y \odot p_{int}^{(c)} - \widehat{Y} \odot p_{int}^{(c)} \right\|_1}{p_{int}^{(c)} \otimes 1} \tag{4}$$

where $\otimes$ denotes the convolutional operation, $\odot$ denotes the pixelwise product operation, $p_{int}^{(c)}$ is the c-th component of intermediate parsing P_{int}, and δ is equal to 0.7.

Perceptual loss. In order to encourage the synthesized photo to have similar features and identity with the target photo, we introduce the perceptual loss. And it can be formulated as

$$L_{per} = \sum_{l \in S} \left\| \varphi^l \left(Y\right) - \varphi^l \left(\hat{Y}\right) \right\|_2 \tag{5}$$

where $\varphi^l(\cdot)$ represents the output features of the l layer in VGG-19 model [17] and S is the set of Conv1-1, Conv3-4, Conv4-1 layers we selected above.

Full objective in two stages.

The generator G_1 and the discriminators D_1, D_{par} in stage-I can be optimized by the following formulation:

$$\left(G_1^*, D_1^*, D_{par}^*\right) = arg \min_{G_1} \max_{D_1, D_{par}} L_{adv1} + \lambda L_{cmp} + \gamma L_{per} + \mu L_{PM} \tag{6}$$

The generator G_2 and the discriminator D_2 in stage-II can be optimized by the following formulation:

$$\left(G_2^*, D_2^*\right) = arg \min_{G_2} \max_{D_2} L_{adv2} + \lambda L_{cmp} + \gamma L_{per} \tag{7}$$

where the λ, γ and μ in Eq.6 and Eq.7 are set to 10, 5, and 10, respectively.

3. EXPERIMENTS

3.1. Settings

Datasets. The CUHK Face Sketch (CUFS) dataset [18] consists of 606 well-aligned face photo-sketch pairs, including three sub-datasets: the CUHK student dataset (188 pairs), the AR dataset (123 pairs) and the XM2VTS dataset (295 pairs). Following the setting in [9], there are 268 face photo-sketch pairs for training and the rest 338 pairs for testing.

Implementation details. To optimize our network, we follow the training strategy in [9] to train our ISEGAN, following the order of $D_1, D_{par} \to D_2 \to G_1 \to G_2$ to conduct one gradient descent step sequentially. We use the Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. Besides, we set the batch size to 1 and the learning rate to 0.0002. And our ISEGAN is trained on a single NVIDIA 2080Ti GPU for 700 epochs.

Criteria. In this work, we employ four performance metrics as the criteria, including the *Fréchet Inception distance* (FID) [19], the *Feature Similarity Index Metric* (FSIM) [20], the *Learned Perceptual Image Patch Similarity* (LPIPS) [21] and the face recognition accuracy by using the *Null-space Linear Discriminant Analysis* (NLDA) [22]. Furthermore, we utilize three classification networks (i.e. AlexNet, SqueezeNet, VGGNet) to calculate the LPIPS value respectively.

3.2. Comparison with State-of-the-art Methods

In this section, we compare our method with various advanced methods. We use the results or codes of these methods released by their corresponding authors.

Quantitative comparison. As is shown in Table 1. Obviously, our method achieves the best FSIM, the best all three kinds of LPIPS (AlexNet, SqueezeNet, VGGNet), the best face recognition accuracy (NLDA) and the second best

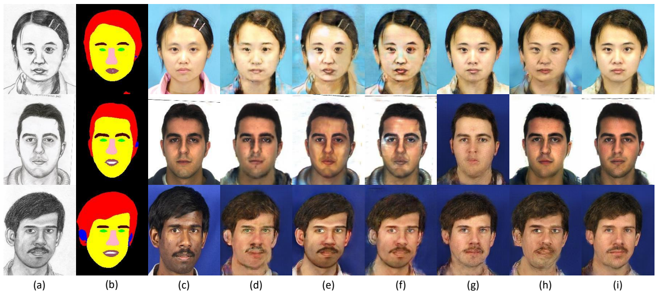

Fig. 3. Examples of synthesized face photos on the CUFS dataset: (a) sketch, (b) sketch parsing, (c) target photo, (d) Pix2pix, (e) CycleGAN, (f) DualGAN, (g) SCA-GAN, (h) CHAN, (i) Ours.

Table 1. 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 <u>underline</u> format, respectively. ↓ indicates the lower is better, and ↑ higher is better.

Methods (Year)	FID↓	FSIM↑	LPIPS↓ (alex)	LPIPS↓ (squeeze)	LPIPS↓ (vgg)	NLDA↑ (%)
Pix2pix [3] (2017)	73.43	0.7678	0.1981	0.1769	0.3256	<u>93.32</u>
CycleGAN [4](2017)	71.81	0.7682	0.1876	0.1685	0.3279	87.29
DualGAN [5] (2017)	70.65	0.7716	0.2112	0.1846	0.3210	92.37
SCA-GAN [9] (2020)	60.72	0.7764	0.1790	<u>0.1539</u>	<u>0.3011</u>	90.61
CHAN [14] (2021)	**41.44**	<u>0.7781</u>	<u>0.1788</u>	0.1553	0.3044	87.79
ISEGAN(ours)	<u>58.30</u>	**0.7802**	**0.1668**	**0.1432**	**0.2895**	**94.02**

FID. Compared with previous state-of-the-art methods, our ISEGAN has a great advantage on performance: our ISEGAN greatly decreases the previous best LPIPS-alex from 0.1788 to 0.1668, LPIPS-squeeze from 0.1539 to 0.1432 and LPIPS-vgg from 0.3011 to 0.2895. Besides, our ISEGAN has a considerable improvement on face recognition accuracy (NLDA) from the previous best 93.32 to 94.02 and increases the previous best FSIM from 0.7781 to 0.7802. In fact, the FID is not suitable for our task [8] because the FID evaluates the ability to generate various classes of object that our task does not concern. To prove the superiority of our method in different aspects, we remain the comparison on FID.

Qualitative comparison. As shown in Fig. 3, our ISEGAN can generate sharp and realistic face photos that have rich facial details and prolific identity features. Moreover, our ISEGAN can effectively solve the color aliasing problem in CycleGAN and DualGAN and the serious deformations and aliasing problems in SCA-GAN and CHAN. Besides, due to the lack of color information in the input sketches, it is normal that the skin color of synthesized photos is not correct, which is a limitation in the sketch-photo synthesis task.

3.3. Ablation Study

In this part, we will conduct abundant ablation experiments to prove the effectiveness of our design choices. The effectiveness of the input sketch parsing has been proved in [9, 12, 15]. In addition, the generated intermediate face parsing participates in the calculation of the compositional loss and the PM loss. As a result, for the rigor of ablation studies, we will conduct experiments to prove the effectiveness of (1) L_{per}; (2) L_{cmp}; (3) L_{PM}; (4) using P_{int} to enhance P_s; (5) SE-GAN. The results of ablation experiments on the CUFS dataset are presented in Table 2 and Fig. 4.

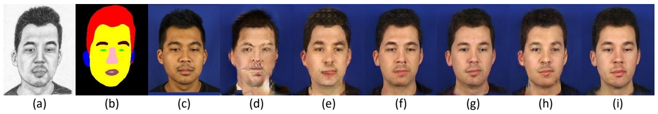

Fig. 4. Ablation experiments of our ISEGAN on the CUFS dataset: (a) sketch, (b) sketch parsing, (c) target photo, (d)-(i) correspond to the settings in Table 2 from top to bottom.

Table 2. Ablation experiments on the CUFS dataset.

Methods	FID↓	FSIM↑	LPIPS↓ (alex)	LPIPS↓ (squeeze)	LPIPS↓ (vgg)	NLDA↑ (%)
Base	85.96	0.7480	0.2624	0.2401	0.4140	78.72
Base + L_{per}	65.04	0.7683	0.2093	0.1826	0.3256	86.84
Base + L_{per} + L_{cmp}	<u>59.53</u>	0.7753	0.1742	0.1510	0.2983	91.70
Base + L_{per} + L_{cmp} + L_{PM}	61.42	0.7775	<u>0.1715</u>	<u>0.1469</u>	<u>0.2940</u>	92.53
full model (stage-I output)	64.34	<u>0.7801</u>	0.1757	0.1565	0.3062	<u>93.30</u>
full model (stage-II output)	**58.30**	**0.7802**	**0.1668**	**0.1432**	**0.2895**	**94.02**

As Fig. 4 (e) shows, the perceptual loss encourages a recognizable face. As Fig. 4 (f) presents, the compositional loss enriches facial details and encourages favorable facial components. As illustrated in Fig. 4 (g), the parsing matching loss obviously alleviates the deformations and aliasing on eyes and mouth regions. Compared with Fig. 4 (g), (i) shows that the enhanced face parsing is significantly effective in solving the deformations and aliasing on eyes and mouth regions. What is more, the noise and artifacts around facial components are greatly reduced. Compared with Fig. 4 (h), (i) demonstrates that the SE-GAN is effective, driving ISEGAN to generate high-quality realistic face photos.

As shown in Table 2, the performance improves gradually with the model becoming complete. The ablation experiments above prove that each design of our ISEGAN is meaningful. What is more, our full model achieves the best performance among all models in ablation experiments.

4. CONCLUSION

In this paper, we propose an intermediate face parsing to enhance the semantic information of the input parsing, which can solve the problems caused by the semantic mismatching. Based on this, an Intermediate Semantic Enhancement Adversarial Network (ISEGAN) model is proposed. Besides, we propose a parsing matching loss to further improve the intermediate face parsing. We conduct extensive comparison experiments on our ISEGAN and the previous state-of-the-art methods. As a result, our ISEGAN achieves the best performance. In the future, we will pay more effort to improve our method to achieve unsupervised face sketch-photo synthesis.

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

[3] P. Isola, J. Zhu, and T. Zhou 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.

[4] J. Zhu, T. Park, and P. Isola 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.

[5] Z. Yi, H. Zhang, and P. Tan 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.

[6] 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.

[7] S. Zhang, R. Ji, and J. Hu et al., "Face sketch synthesis by multidomain adversarial learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 5, pp. 1419–1428, 2019.

[8] M. Zhu, J. Li, and N. Wang et al., "Knowledge distillation for face photo-sketch synthesis," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2020.

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

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

[11] H. Zhang, T. Xu, and H. Li 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.

[12] 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.

[13] T. Park, M. Liu, and T. Wang 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.

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

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

[16] J. Ma, H. Xu, and J. Jiang et al., "Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion," *IEEE Transactions on Image Processing*, vol. 29, pp. 4980–4995, 2020.

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

[18] 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.

[19] Heusel, Martin, and Ramsauer et al., "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, pp. 6626–6637, 2017.

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

[21] Zhang, Richard, and Isola 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.

[22] Li-Fen Chen, Hong-Yuan Mark Liao, and Ming-Tat Ko 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.