LAGAN: Landmark Aided Text to Face Sketch Generation

Wentao Chao, Liang Chang*, Fangfang Xi, Fuqing Duan

Beijing Normal University, Beijing, China
{changliang,xifangfang,fqduan}@bnu.edu.cn
chaowentao@mail.bnu.edu.cn

Abstract. Face sketch is a concise representation of the human face, and it has a variety of applications in criminal investigation, biometrics, and social entertainment. It is well known that facial attribute is an underlying representation of the facial description. However, generating vivid face sketches, especially sketches with rich details, from given facial attributes text is still a challenging task as the text information is limited. Existing work synthetic face sketch is not realistic, especially the facial areas are not natural enough, even distorted. We aim to relieve the situation by introducing face prior knowledge, such as landmarks. This paper proposes a method, called LAGAN, that Landmark Aided Text to Face Sketch Generation. Specifically, we design a novel scale translation-invariant similarity loss based on the facial landmarks. It can measure the mutual similarity between real sketch and synthetic sketch and also measure the self similarity based on the symmetry of face attributes. Further to counter data deficiency, we construct a novel facial attribute text to sketch dataset called TextCUFSF with CUFSF face sketch dataset. Each sketch has 4 manual annotations. Qualitative and quantitative experiments demonstrate the effectiveness of our proposed method for sketch synthesis with attribute text. The code and data are available: https://github.com/chaowentao/LAGAN.

Keywords: Face Sketch · Text to Sketch · Generative Adversarial Nets

1 Introduction

Face sketches play an important role in forensic investigation and biometrics [23, 34]. Face sketches drawn according to the witness' depictions provide an important reference for forensic investigation, especially when the face photos of the suspect are not available. However, sketching by well-trained and skilled artists is labor-intensive. So, it is desired to generate face sketches from face attribute depictions.

The existing face sketch synthesis [8, 23, 34, 44, 46] mostly focuses on generating face sketches based on photos. Wang *et al.* [32] give a comprehensive

* Corresponding Author

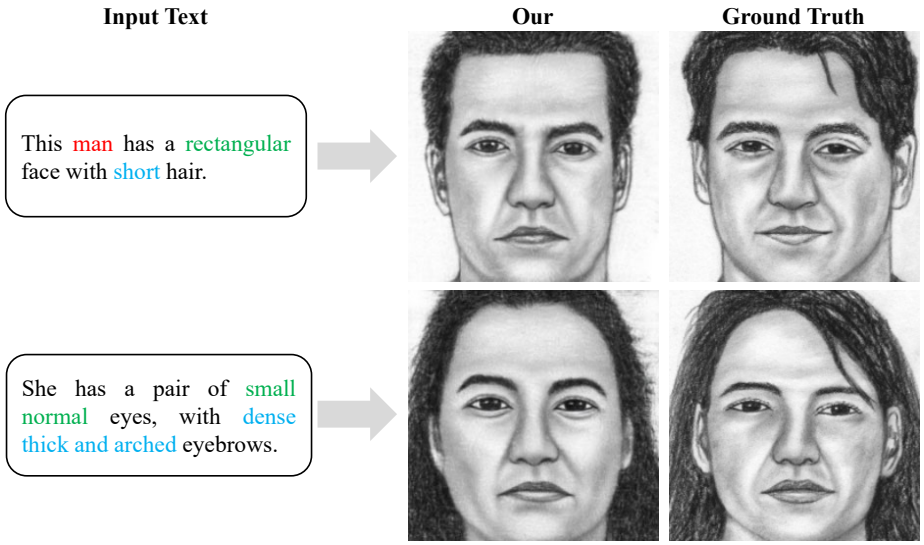

Fig. 1. Illustration of face sketches synthesis from face attribute text. From left to right: Input text, Face sketches with our LAGAN and Ground Truth.

survey on face hallucination and sketch synthesis. Various photo-sketch synthesis methods have been proposed, including the traditional methods of subspace representation [27], sparse representation [2], and Markov random field [34]. Recent work of face sketch synthesis is mainly built upon deep neural network or variable autoencoder (VAE) [9]. However, in a real scenario, even a single photo may not be available for sketch synthesis.

Recently, great progress has been made on synthesizing natural images form text [7,13,17,18,28,40,42,43]. These works generally use generate natural images with text descriptions, and can not meet the requirement of face sketch synthesis with rich facial features, because these methods overlook the key details of face structure and textures. The main difficulties of face sketch synthesis from attribute text lie in two aspects: 1) Face sketch contains rich detail information, and it is necessary to recover more facial details in the face sketch synthesis task from attribute text compared to previous work on synthesizing natural images from the text; 2) No text-to-face sketch database is available. Due to the ambiguity and diversity of sketches, an important component of the database, i.e. the attributes for sketches, is not easy to collect; Moreover, existing sketch databases are relatively smaller compared with the off-the-shelf image and graph databases, such as ImageNet [4] and ShapeNet [1].

Face landmarks are important feature points of each part of the face, usually contour points and corners, including eyebrows, eyes, nose, mouth, face contour. Considering geometric constraints inherent in human faces, we propose a method that landmark aided text to face sketch generation, named LAGAN.

In order to measure the similarity of landmarks, we propose a new similarity loss, which is invariant of scale and translation. Then, the mutual similarity of landmark between real sketch and the synthetic sketch is weighed by similarity loss. Considering the symmetry of the face, we also measure the self similarity of landmark between synthetic sketch and synthetic sketch flipped horizontally. In order to address the data deficiency issue, we built a text-to-sketch dataset named TextCUFSF. It contains 1,139 sketches, and then we annotate manually each sketch with 4 facial attribute text. Benefiting from our dataset, our LAGAN can successfully generate face sketches with rich details. Some examples are shown in Figure 1.

Our contribution lies in the following aspects: 1) We propose the text to face sketch generation method aided by facial landmarks. Given an attribute text, our method can generate a high-quality sketch corresponding to the attribute text; 2) We propose a novel similarity loss for facial landmarks, which is used to measure the mutual similarity and self similarity; 3) We construct a new facial attribute text to the sketch dataset named TextCUFSF based on the CUFSF face sketch dataset.

2 Related Work

Attribute to Image Synthesis. Attribute-to-image aims to generate synthetic images based on attribute descriptions. Attribute2Image [38] develops a layered generative model with disentangled latent variables, which conditioned image generation from visual attributes using VAE. Attribute2Sketch2Face [5], which is based on a combination of a deep conditional variational autoencoder and generative adversarial networks. Han *et al.* [8] present a deep multi-task learning approach to jointly estimate multiple heterogeneous attributes from a single face image.

Text to Natural Image Synthesis. Prior art methods to generate images using text include [7, 13, 17, 18, 28, 42, 43]. These methods generate images using text as input. Reed *et al.* [17] used conditional GAN to generate low-resolution natural images of 64×64 pixels with a text description. The subsequent work GAWWN [18] can generate higher resolution images of 128×128 pixels with additional annotations, such as the bounding box or part of the key points. StackGAN [43] is mostly related to our problem, which uses two stacked GAN networks to get the high-resolution natural images of 256×256 pixels using text as input. StackGAN++ [42] exploits multiple generators and discriminators of a tree structure. Yuan *et al.* [40] propose symmetrical distillation networks to address the problem of heterogeneous and homogeneous gaps in the text-to-image synthesis task. Gorti *et al.* [7] introduce a text-to-image translation GAN and image-to-text translation GAN by enforcing cycle consistency. AttnGAN [37] designs the attention module to fusing the feature of text and images. ControlGAN [12] further leverages the spatial and channel-wise attention module to generate better images. These methods [12, 37, 42, 43] area cascaded in multiple stages to gradually improve the resolution of synthetic images, while some

recent methods [13, 28] can directly synthesize the high-resolution images. DF-GAN [28] designs the new deep fusion block to fuse the information of text and image. SSA-GAN [13] introduces the semantic mask to guide the process of text and image. Compared to text to natural image synthesis, face sketch generation from attribute text is more challenging due to the fact that more appearance details are in face sketches. These methods use different GAN models to synthesize natural images from text and overlook the structured priors of the output. Compared to text to natural image synthesis, face sketch generation from attribute text is more challenging due to the key details of face structure and texture.

Face Photo-to-Sketch and Sketch-to-Photo Synthesis. These methods aim to synthesize face sketches using face photos, which is called "Face photo-sketch synthesis" [30, 31, 34]. Wang *et al.* [32] gives a comprehensive survey on face hallucination and sketch synthesis. Various photo-sketch synthesis methods have been proposed, including the traditional methods of subspace representation [27], sparse representation [2], and Markov random field [34]. Recent work of face sketch synthesis is mainly built upon a deep neural network or variable autoencoder (VAE) [9]. Concerning photo synthesis from a given sketch, Sangkloy *et al.* [20] propose a deep adversarial image synthesis architecture, named Scribbler. It is conditioned on sketched boundaries and sparse color strokes to generate realistic photos, including face photos. The feed-forward architecture is fast and interactive. Wang *et al.* [44] propose a photo-face sketch synthesis framework consisting of a coarse stage and a refine stage. In the coarse stage, a coarse face sketch is synthesized. In the refine stage, a probabilistic graphic model starts by erasing the noise of the coarse synthesized sketches and then obtains the details on the facial components. DualGAN [39] enables both sketch-to-photo and photo-to-sketch generation from unlabeled data. Compared to photo-to-sketch and sketch-to-photo synthesis, attribute text to sketch synthesis addressed in this work is more challenging due to the large domain difference between a sketch and attribute text.

Generative Adversarial Networks and its Applications. Generative Adversarial Networks (GAN) [6] shows promising performance for generating photo-realistic images. Various GANs have been proposed to manipulate face images. On face image editing, Choi *et al.* [3] propose StarGAN, which performs image-to-image conversion on multiple domains by utilizing a mask vector method to control all available domain labels and is effective in facial attribute transfer as well as facial expression synthesis tasks. Shen *et al.* [21] address the problem of face attribute manipulation by modifying a face image according to a given attribute value. On face image synthesis, M-AAE [24] combines the VAE and GAN to generate photo-realistic images. Face recognition loss, cycle consistency loss, and facial mask loss are integrated with the network. G2-GAN [22] employs facial landmarks as a controllable condition to guide facial texture synthesis under specific expressions.

In this work, we propose a method to generate face sketches with face attribute text. Considering the geometry of face, we also define the mutual similarity and self similarity based on the face landmarks.

3 TextCUFSF Dataset

Sketch Dataset. We build our text-sketch dataset named TextCUFSF based the CUFSF [33, 45] dataset. CUFSF includes 1,139 photos from the FERET dataset and detailed face sketches drawn by an artist. The main reason that we use CUFSF for annotation is that CUFSF is one of the largest public sketch datasets with a wide range. We selected 1,000 pairs of sketches as train sets and the rest as test sets. We align the face sketches using face landmarks such as eye centers and nose points, which are extracted with the SDM method [36].

Face Attributes. We design our facial attribute set inspired by a seminal criminal appearance study [10,29]. In criminal appearance studies, facial appearance characteristics can be divided into eleven categories, namely hairstyle, head shape, wrinkles, eyebrows, eyes, nose, ears, mouth, beard, hat, and other features. We choose eight key features from them and add gender, and glasses features (refer to Table 1 for details). We do not take features such as expressions and wrinkles into consideration, since they are difficult to accurately characterize the face sketches. Then, we choose representative words for each feature based on the theory in criminal appearance studies.

Table 1. Face Attribute Description.

Face Components	Attribute Description
Face	diamond, heart, triangle, round, inverted triangle, oval, square, rectangular
Eyebrows	dense, sparse, thick, thin, flat, arched, up, down
Eyes	big, normal, short, wide, narrow,
Hair	long, medium, short
Nose	big, medium, small
Mouth	thick, thin, wide, narrow
Ears	small, normal, big
Glasses	has, hasn't
Beard	has, hasn't
Gender	man, woman

Specifically, the face has eight types. The eyebrow is characterized according to density, thickness and shape. The nose is characterized according to the height and the size of the nostrils. The lip is characterized in terms of the thickness and width. The ear is represented by the size. Beard and glasses are taken into account as well.

Face Attribute Text Annotation. We invite 50 annotators to annotate the 1,139 sketches in the CUSFS dataset. Each sketch has 4 captions, as shown in Figure 2. In order to enforce the quality of data annotation, we prepare well-annotated attribute sketch pairs for the qualification tests. The qualification tests are pre-conducted for every annotator using three different sketches, and those who pass the tests are qualified for the annotation. For the annotated results, we then manually review and correct them to ensure the annotation accuracy.

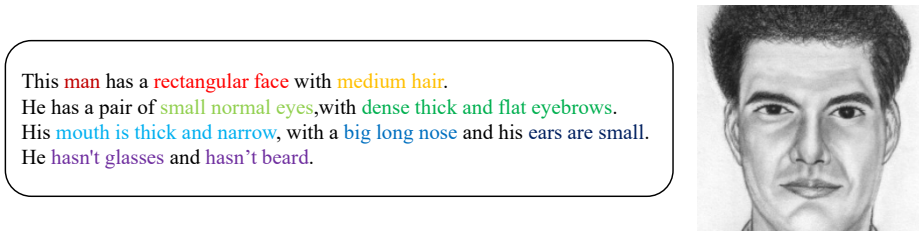

Fig. 2. Examples of face sketches and face attribute text. From left to right: Face attribute text, Faces sketches drawn by artists.

4 Method

In this section, we propose the LAGAN network to generate the sketch from our TextCUFSF dataset. Our method can synthesize the 256×256 high-resolution face sketches based on the given attribute text. Firstly, we introduce the pipeline of our LAGAN. Then, the loss functions used are described. Finally, we present the experimental implementation details.

4.1 Network Structure

Figure 3 (a) shows the structure of our LAGAN. Inspired by AttnGAN [37], we adopt a similar multi-stage mode to synthesize face sketches. For the sake of presentation, some intermediate results are not presented. We first convert the input facial attribute text to text embedding through a text encoder, then use a face sketch synthesis model with various loss functions for sketch synthesis. Finally, we use a discriminator to distinguish between true and false synthetic sketches and true and false matches between attribute text and sketches.

Generator We can encode the attribute text t into a sentence embedding s and word embedding w by text encoder, which is a pre-trained bidirectional LSTM. The sentence embedding s is augmented by Conditioning augmentation (CA) [43]

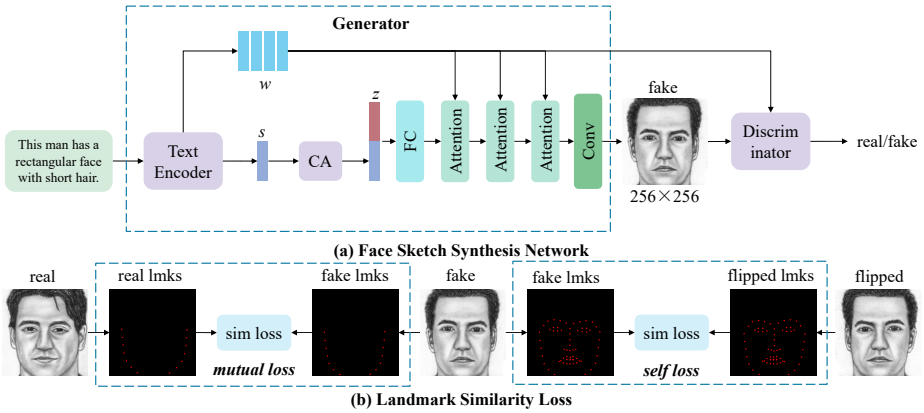

Fig. 3. The pipeline of our LAGAN. (a) Face Sketch Synthesis Network. It consists of a generator and a discriminator. The input of the generator is the facial attribute text and random Gaussian noises, the output is the synthesized face sketches of 256×256. The function of the discriminator is used in both distinguishing between the real and false synthesized sketch as well as the true and false matching of attribute text and sketch. (b) Landmark Similarity Loss. We show schematic diagrams of mutual loss and self Loss.

module. Augmented sentence embedding s and random noise (e.g. Gaussian Distribution) z are concatenated and through FC layers as input to the attention module [12]. Specifically, the function of the attention module calculates the spatial and channel-wise correlation between image features and word embedding. After three attention modules, we can obtain the final synthesized face sketches $\hat{I}$ of 256×256 pixels.

Discriminator The input of the discriminator has two parts that text embedding and face sketches. Further, we extract the image feature $\hat{\eta}$ of the synthesized face sketch $\hat{I}$ by GoogleNet-based [25] image encoder. The discriminator is designed to compute the word-level text-sketch correlation between word embedding w and image feature $\hat{\eta}$, and also to distinguish the quality of synthetic sketches, real or false.

4.2 Loss Functions

In this section, we introduce landmark similarity loss and the GAN loss, then summarize the total generator loss and discriminator loss.

Landmark Similarity Loss We find previous methods [12,13,28,37,43] are mainly text-to-image task, such as generate birds and flowers. Such objects' structures compared with face sketches are simpler and the information in the text is limited. When we apply these methods in face sketches, synthesized results are not

satisfactory, which are not natural and some facial areas are not reasonable, even distorted. So we aim to add face prior knowledge to relieve these problems, such as facing landmark constrain. Based on Riemannian distance, we design a novel invariant landmark similarity loss. The specific operation process is shown in Algorithm 1. Specifically, the algorithm input are two lists of 2D face landmarks, e.g. l_a and l_b. The main steps of the algorithm are to subtract l_{ma}, l_{mb} and regularize the mean values of each landmark l_{na}, l_{nb}, then generate the similarity matrix m_{ab}, calculate the trace of the matrix t_{ab}, and finally obtain the similarity sim through $arccos$ operation. Trace is the sum of the main diagonal elements of the matrix, reflecting the degree of similarity between the same index landmarks. Then, mutual loss $\mathcal{L}_{mutual}$ and self loss $\mathcal{L}_{self}$ are generated, as shown in Figure 3(b).

Algorithm 1 Face Landmark Cosine Similarity

Input: l_a: A face landmarks; l_b: B face landmarks;
Output: sim: cosine similarity;

1: De-mean, subtract mean separately
$l_{ma} = l_a - mean(l_a)$; $l_{mb} = l_b - mean(l_b)$;
2: Normalization, divide their norm
$l_{na} = l_{ma} / \|l_{ma}\|$; $l_{nb} = l_{mb} / \|l_{mb}\|$;
3: Calculate similarity matrix
$m_{ab} = l_{na}^T \times l_{nb}$;
4: Calculate absolute value of matrix trace
$t_{ab} = |trace(m_{ab})|$;
5: Calculate cosine similarity
$sim = 1 - \arccos(min(t_{ab}, 1))$;

Mutual Loss In order to obtain realistic facial areas corresponding to the given attribute text, the mutual loss $\mathcal{L}_{mutual}$ is proposed to calculate the similarity between the landmarks of real sketches I and synthesized face sketches $\hat{I}$. We use the portion of landmarks, which show the area of attribute text, e.g. face, eyes, or mouth. The loss function of $\mathcal{L}_{mutual}$ is as follows:

$$\mathcal{L}_{mutual} = similarity\left(I, \hat{I}\right) \tag{1}$$

Self Loss We also find the symmetry of face prior knowledge. Specifically, areas of the left face attribute are similar to the right. Therefore, self loss $\mathcal{L}_{self}$ is used to calculate the similarity synthesized sketches $\hat{I}$ and corresponding flipped synthesized sketches $\hat{I}_{flip}$. add symmetry constraints. When calculating self loss, all facial landmarks are considered. The loss $\mathcal{L}_{self}$ can be defined as follows:

$$\mathcal{L}_{self} = similarity\left(I, \hat{I}_{flip}\right) \tag{2}$$

GAN Loss Our method is based on conditional generative adversarial network (CGAN) [15], which learns a mapping $y = G(c, z)$ from input c and random noise

vector z to y. The goal of discriminator D is to correctly distinguish between the real sketches and the synthesized sketches by generator G, while the goal of generator G is to generate synthesized sketches that can fool the discriminator D. The objective function of GAN in this work is:

$$\min_G \max_D V(D,G) = \mathbb{E}_{I\sim p_{data}}[D(I,t)] + \mathbb{E}_{z\sim p_z}[1 - D(G(z,t),t)] \tag{3}$$

where I is a real face sketch from the true data distribution p_{data}, z is a noise vector sampled from distribution p_z, and t is a given facial attribute text.

The loss function of D is as follows:

$$\mathcal{L}_{GAN}(D) = \mathbb{E}_{I\sim p_{data}}[D(I,t)] + \mathbb{E}_{z\sim p_z}[1 - D(G(z,t),t)] \tag{4}$$

The loss function of G is as follows:

$$\mathcal{L}_{GAN}(G) = \mathbb{E}_{z\sim p_z}[1 - D(G(z,t),t)] \tag{5}$$

Generator Loss Inspired by these work [12, 37], we also use perceptual loss and DAMSM loss. So the total loss of the generator $\mathcal{L}_G$ contains GAN loss, mutual loss, self loss, perceptual loss and DAMSM loss, which is defined as the follows:

$$\mathcal{L}_G = \mathcal{L}_{GAN}(G) + \lambda_1\mathcal{L}_{DAMSM} + \lambda_2\mathcal{L}_{per} + \lambda_3\mathcal{L}_{mutual} + \lambda_4\mathcal{L}_{self} \tag{6}$$

Discriminator Loss The objective function of discriminators $\mathcal{L}_D$ is mainly the GAN loss defined as Equation (4).

$$\mathcal{L}_D = \mathcal{L}_{GAN}(D) \tag{7}$$

4.3 Implementation Details

The structure of the model is shown in Fig. 3. Our experiments are performed on our TextCUFSF dataset, with 1,000 pairs of text sketches for training and 100 pairs for testing. We use a horizontal flip for data augmentation. The dimension of the encoded text attribute is 100. In our experiment, hyperparameters are set as: $\lambda_1 = 5$, $\lambda_1 = 1$, $\lambda_3 = 100$ and $\lambda_4 = 100$. Dlib module is employed to extract 68 facial landmarks. Our LAGAN is implemented by Pytorch and trained with a learning rate of 0.002. We set 600 epochs for training, with a batchsize of 8. The generator and discriminator have trained alternatively. The training time is about 15 hours on a 1080 Ti GPU.

5 Experiments

In order to evaluate our method, we conduct comparison experiments with two related state-of-the-art methods and ablation study on the TextCUFSF dataset.

5.1 Evaluation Metrics

In this work, we mainly use two standard measures, i.e. FID (Fréchet Inception Distance) [14] and SSIM (Structure Similarity Index Measure) [35] for performance evaluations.

FID FID (Fréchet Inception Distance) [14] can capture the similarity of generated images to real ones, it has better performance than the Inception Score [19]. It is consistent with human visual system, and is often used to evaluate the quality of results from GAN [11, 16, 41]. FID is calculated by the Fréchet distance between two Gaussian's fitted to the feature representations of the Inception network.

SSIM SSIM (Structure Similarity Index Measure) [35] indicates the local structural similarity between the image segmentation result and the reference image. It is widely used to evaluate the similarities between synthesized face sketches and real face sketches [26].

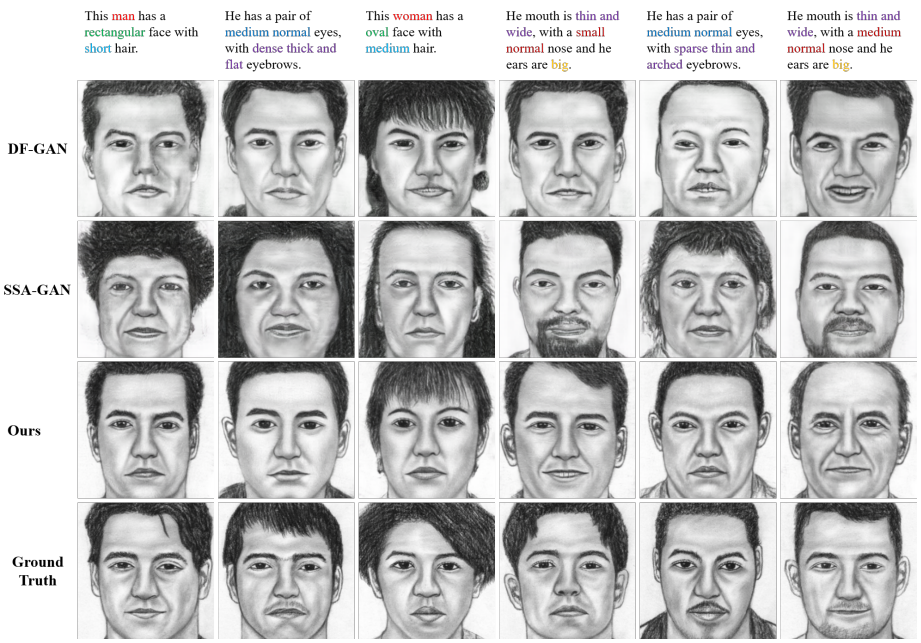

Fig. 4. Example of synthesized face sketches by DF-GAN [28], SSA-GAN [13], our method given face attribute description text and Ground Truth. The resolution of face sketches is 256×256.

5.2 Comparison with the State-of-the-art Methods

We compare our results with the state-of-the-art methods of DF-GAN [28] and SSA-GAN [13] on our TextCUFSF dataset. As shown in Fig. 4, all the three models can produce the face sketches of 256×256 pixels. In terms of quality, some of the results of DF-GAN are ambiguous and coarse, such as the areas of face shape, ears, and eyes on the first, third, and fifth sketches in Fig. 4, respectively. The quality of sketches generated by SSA-GAN is better, while some attributes are not complete, such as the beard area. Our method can generate high-resolution sketches with complete facial areas. Considering the consistency, SSA-GAN is not well consistent with the input attribute text, such as gender attribute. DF-GAN is more consistent with the input attribute text than SSA-GAN, while some attributes are of low quality. Our method can achieve the best synthesized sketches that are more consistent and high-quality with the input attribute text.

In addition to two standard evaluation indicators, we also compare the MS (Mutual Similarity) and SS (Self Similarity) to validate the similarity of face landmarks. Table 2 illustrates the quantitative evaluations. Our LAGAN performs best in all evaluation metrics. The FID of our method is 29.60, which is 19.16 higher than DF-GAN, and 15.83 higher than SSA-GAN. The PSNR and MSSIM of our method are 13.01% and 32.81% higher than StackGAN, respectively. Our method also achieved 0.03-0.05 improvement in SSIM index compared with two other methods. We also observe that our method can achieve best result when calculating the similarity of face landmarks.

Table 2. Quantitative comparison of our LAGAN with state-of-the-art methods, including FID, SSIM, MS (Mutual Similarity), SS (Self Similarity) , MR (Match Rank) and QR (Quality Rank).

Method	FID↓	SSIM ↑	MS ↑	SS ↑	MR ↓	QR ↓
DF-GAN [28]	48.76	0.3005	0.88	0.88	2.80	2.78
SSA-GAN [13]	45.43	0.2745	0.89	0.91	1.98	1.86
Ours	**29.60**	**0.3311**	**0.93**	**0.94**	**1.22**	**1.36**

User Study Although the above metrics correlate well with human perception on visual quality [19], it cannot be used to measure the matching of sketch with attribute text and the quality of sketch. We design user study for matching and quality evaluations. The matching evaluation aims to rank the consistence between the synthesized sketch and input attribute text, 1 represents the highest priority, and 3 represents the lowest priority. The quality evaluation is similar to the matching evaluation except without providing the text description. We randomly invited 48 people to perform user study of matching and quality evaluations. Table 2 shows the MR (Match Rank) and QR (Quality Rank) results, our method gets the highest priority. The consistency and quality of the sketch synthesized by our method are more recognized by people.

Table 3. Comparison using different combinations of loss functions and different hyper-parameters with the our LAGAN.

Mutual Loss	Self Loss	FID↓	SSIM ↑
		34.13	0.3108
	✓	36.78	**0.3321**
✓		35.97	0.3315
✓	✓	**29.60**	0.3311
λ_3=1	λ_4=1	31.26	0.3293
λ_3=10	λ_4=10	38.57	0.3356
λ_3=10	λ_4=100	33.79	0.3262
λ_3=100	λ_4=10	32.22	**0.3372**
λ_3=100	λ_4=100	**29.60**	0.3311
λ_3=300	λ_4=300	32.26	0.3365
λ_3=1000	λ_4=1000	33.85	0.3363

Ablation Study Firstly, We analyze the effect of face landmark similarity on the performance of sketch synthesis. The quantitative evaluation for our LAGAN with different combinations of loss functions are shown in Table 3. We observe that our model combined with mutual loss and self loss can get the lowest value of FID and comparable results of SSIM. Then, we carry out detailed experiments to find the optimal hyper-parameters of mutual loss and self loss. When $\lambda_3 = 100$ and $\lambda_4 = 100$, our LAGAN can achieve better performance. Finally, we also compare the results of DF-GAN and SSA-GAN with our face landmark similarity loss. Table 4 shows that SSIM have been improved in different models using our loss, which validate the generalization ability of our loss.

Table 4. The performance of face landmark similarity loss on different models

Model	Mutual Loss	Self Loss	SSIM↑
DF-GAN [28]			0.3005
	✓	✓	**0.3240**(+0.02)
SSA-GAN [13]			0.2745
	✓	✓	**0.2855**(+0.01)
LAGAN			0.3108
	✓	✓	**0.3311**(+0.02)

6 Conclusion

In this work, we provide a LAGAN method to generate face sketches based on attribute text. We also design a novel landmark similarity loss, which is invariant of scale and translation. Based on the similarity loss, we further measure the mutual similarity and self similarity. We construct the first attribute text

to face sketch dataset, called TextCUFSF. Qualitative and quantitative experiments show that our method can obtain high-quality synthesized face sketches from attribute text. Our method also has some limitations regarding beard and glasses attributes, we analyze the reason due to the training samples of the related attributes are too few. By collecting more samples, this situation can be alleviated. In the future, we intend to support text-based sketch editing for interactive, dynamic synthesis of high-quality sketches to meet artist expectations.

Acknowledgement. This work was supported by the National Key Research and Development Program of China under grant No. 2019YFC1521104, Natural Science Foundation of China (61772050, 62172247).

References

1. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015)
2. Chang, L., Zhou, M., Han, Y., Deng, X.: Face sketch synthesis via sparse representation. In: ICPR. pp. 2146–2149 (2010)
3. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: CVPR. pp. 8789–8797 (2018)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009)
5. Di, X., Patel, V.M.: Face synthesis from visual attributes via sketch using conditional vaes and gans. arXiv preprint arXiv:1801.00077 (2017)
6. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NeurIPS. pp. 2672–2680 (2014)
7. Gorti, S.K., Ma, J.: Text-to-image-to-text translation using cycle consistent adversarial networks. arXiv preprint arXiv:1808.04538 (2018)
8. Han, H., Jain, A.K., Wang, F., Shan, S., Chen, X.: Heterogeneous face attribute estimation: A deep multi-task learning approach. TPAMI **40**(11), 2597–2609 (2017)
9. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
10. Klare, B.F., Klum, S., Klontz, J.C., Taborsky, E., Akgul, T., Jain, A.K.: Suspect identification based on descriptive facial attributes. In: IJCB. pp. 1–8 (2014)
11. Kurach, K., Lucic, M., Zhai, X., Michalski, M., Gelly, S.: The GAN landscape: Losses, architectures, regularization, and normalization. CoRR **abs/1807.04720** (2018)
12. Li, B., Qi, X., Lukasiewicz, T., Torr, P.H.: Controllable text-to-image generation. In: NeurIPS. pp. 2065–2075 (2019)
13. Liao, W., Hu, K., Yang, M.Y., Rosenhahn, B.: Text to image generation with semantic-spatial aware gan. arXiv preprint arXiv:2104.00567 (2021)
14. Lucic, M., Kurach, K., Michalski, M., Gelly, S., Bousquet, O.: Are gans created equal? a large-scale study. arXiv preprint arXiv:1711.10337 (2017)
15. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)

16. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. arXiv preprint arXiv:1802.05957 (2018)
17. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: ICML. pp. 1060–1069 (2016)
18. Reed, S.E., Akata, Z., Mohan, S., Tenka, S., Schiele, B., Lee, H.: Learning what and where to draw. In: NeurIPS. pp. 217–225 (2016)
19. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: NeurIPS. pp. 2234–2242 (2016)
20. Sangkloy, P., Lu, J., Fang, C., Yu, F., Hays, J.: Scribbler: Controlling deep image synthesis with sketch and color. In: CVPR. pp. 6836–6845 (2017)
21. Shen, W., Liu, R.: Learning residual images for face attribute manipulation. In: CVPR. pp. 1225–1233 (2017)
22. Song, L., Lu, Z., He, R., Sun, Z., Tan, T.: Geometry guided adversarial facial expression synthesis. arXiv preprint arXiv:1712.03474 (2017)
23. Song, Y., Bao, L., Yang, Q., Yang, M.H.: Real-time exemplar-based face sketch synthesis. In: ECCV. pp. 800–813 (2014)
24. Sun, R., Huang, C., Shi, J., Ma, L.: Mask-aware photorealistic face attribute manipulation. arXiv preprint arXiv:1804.08882 (2018)
25. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR. pp. 1–9 (2015)
26. Tan, Y., Tang, L., Wang, X.: An improved criminisi inpainting algorithm based on sketch image. Journal of Computational and Theoretical Nanoscience **14**(8), 3851–3860 (2017)
27. Tang, X., Wang, X.: Face sketch recognition. TCSVT **14**(1), 50–57 (2004)
28. Tao, M., Tang, H., Wu, S., Sebe, N., Jing, X.Y., Wu, F., Bao, B.: Df-gan: A simple and effective baseline for text-to-image synthesis. arXiv preprint arXiv:2008.05865 (2020)
29. Tome, P., Vera-Rodriguez, R., Fierrez, J., Ortega-Garcia, J.: Facial soft biometric features for forensic face recognition. Forensic science international **257**, 271–284 (2015)
30. Wang, N., Gao, X., Sun, L., Li, J.: Bayesian face sketch synthesis. TIP **26**(3), 1264–1274 (2017)
31. Wang, N., Li, J., Sun, L., Song, B., Gao, X.: Training-free synthesized face sketch recognition using image quality assessment metrics. arXiv preprint arXiv:1603.07823 (2016)
32. Wang, N., Tao, D., Gao, X., Li, X., Li, J.: A comprehensive survey to face hallucination. IJCV **106**(1), 9–30 (2014)
33. Wang, X., Tang, X.: Face photo-sketch synthesis and recognition. TPAMI **31**(11), 1955–1967 (2008)
34. Wang, X., Tang, X.: Face photo-sketch synthesis and recognition. TPAMI **31**(11), 1955–67 (2009)
35. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. TIP **13**(4), 600–612 (2004)
36. Xiong, X., Torre, F.D.L.: Supervised descent method and its applications to face alignment. In: CVPR. pp. 532–539 (2013)
37. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X.: Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In: CVPR. pp. 1316–1324 (2018)
38. Yan, X., Yang, J., Sohn, K., Lee, H.: Attribute2image: Conditional image generation from visual attributes. In: ECCV. pp. 776–791 (2016)

39. Yi, Z., Zhang, H., Tan, P., Gong, M.: Dualgan: Unsupervised dual learning for image-to-image translation. In: ICCV. pp. 2868–2876 (2017)
40. Yuan, M., Peng, Y.: Text-to-image synthesis via symmetrical distillation networks. arXiv preprint arXiv:1808.06801 (2018)
41. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. In: ICML. pp. 7354–7363. PMLR (2019)
42. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.: Stackgan++: Realistic image synthesis with stacked generative adversarial networks. arXiv preprint arXiv:1710.10916 (2017)
43. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: ICCV. pp. 5907–5915 (2017)
44. Zhang, M., Wang, N., Li, Y., Wang, R., Gao, X.: Face sketch synthesis from coarse to fine. In: AAAI. pp. 7558–7565 (2018)
45. Zhang, W., Wang, X., Tang, X.: Coupled information-theoretic encoding for face photo-sketch recognition. In: CVPR. pp. 513–520. IEEE (2011)
46. Zou, C., Yu, Q., Du, R., Mo, H., Song, Y., Xiang, T., Gao, C., Chen, B., Zhang, H., et al.: Sketchyscene: Richly-annotated scene sketches. In: ECCV. pp. 438–454 (2018)