Robust Face Sketch Synthesis via Generative Adversarial Fusion of Priors and Parametric Sigmoid

Shengchuan Zhang[1,2], **Rongrong Ji**[1,2*], **Jie Hu**[1,2], **Yue Gao**[3], **Chia-Wen Lin**[4]
[1] Fujian Key Laboratory of Sensing and Computing for Smart City, Xiamen University
[2] School of Information Science and Engineering, Xiamen University
[3] School of Software, Tsinghua University
[4] Department of Electrical Engineering, National Tsing Hua University

Abstract

Despite the extensive progress in face sketch synthesis, existing methods are mostly workable under constrained conditions, such as fixed illumination, pose, background and ethnic origin that are hardly to control in real-world scenarios. The key issue lies in the difficulty to use data under fixed conditions to train a model against imaging variations. In this paper, we propose a novel generative adversarial network termed *p*GAN, which can generate face sketches efficiently using training data under fixed conditions and handle the aforementioned uncontrolled conditions. In *p*GAN, we embed key photo priors into the process of synthesis and design a parametric sigmoid activation function for compensating illumination variations. Compared to the existing methods, we quantitatively demonstrate that the proposed method can work well on face photos in the wild.

1 Introduction

Face sketch synthesis [Wang *et al.*, 2014; Isola *et al.*, 2016] has attracted extensive research focus with broad application prospects ranging from digital entertainment to law enforcement. Towards high-quality sketch synthesis, various approaches [Liu *et al.*, 2005; Wang and Tang, 2009; Zhou *et al.*, 2012; Wang *et al.*, 2013; Song *et al.*, 2014; Zhang *et al.*, 2015a; Wang *et al.*, 2017] have been proposed, which suit well for controlled conditions. These works can be subdivided into two categories, *i.e.*, data-driven methods [Liu *et al.*, 2005; Wang and Tang, 2009; Zhou *et al.*, 2012; Wang *et al.*, 2013; Song *et al.*, 2014] and model-based methods [Zhang *et al.*, 2015a; Wang *et al.*, 2017]. Data-driven methods are consisted of two steps: neighbor selection and weighted reconstruction, whereas model-based methods usually include clustering and regression phases. Recent advances in model-based methods advocating deep learning [Isola *et al.*, 2016; Zhang *et al.*, 2015a] have made great success. However, either data-driven or model-based methods still fail to synthesize face sketches in the wild. In such circumstances, the testing photos are taken under different conditions from the training photos, containing variations of illumination, pose, background and ethnic origin, as shown in Figure 1.

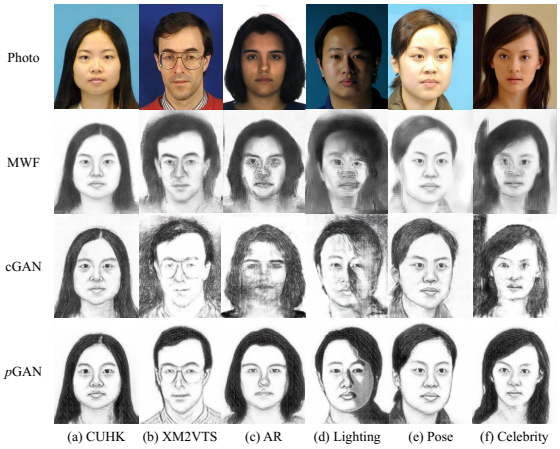

Figure 1: Examples of face sketches generated from face photos taken *in the wild*, *i.e.*, under uncontrolled conditions. We only select 88 subjects in the CUHK student dataset for training. The sketches generated by data-driven method (MWF), model-based method (cGAN) and our method (*p*GAN) are listed in the second, third and fourth rows, respectively. (a)-(f) are the results of different testing photos from (a) the CUHK student dataset, (b) the XM2VTS dataset, (c) the AR dataset, (d) the lighting variation set, (e) the pose variation set and (f) celebrity photos obtained from the Web. Our method achieves satisfactory results under such uncontrolled conditions.

Although there are some works [Zhang *et al.*, 2010; Zhang *et al.*, 2015b; Peng *et al.*, 2016; Zhang *et al.*, 2017; Song *et al.*, 2017; Zhu *et al.*, 2017b] paying attention to this challenging problem, these methods only deal with certain aspects, such as the lighting and pose variations [Zhang *et al.*, 2010], the background and non-facial factors [Zhang *et al.*, 2015b] or the lighting variations and clutter background [Peng *et al.*, 2016]. Most existing methods simulate the process of sketch synthesis by using low level features, which is hardly to be generalized to uncontrolled conditions. In con-

*Corresponding author: Rongrong Ji (rrji@xmu.edu.cn)

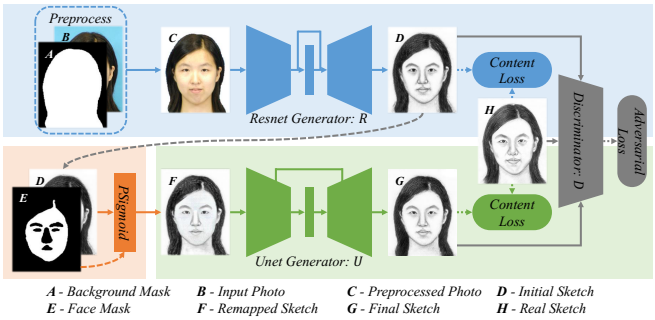

Figure 2: Framework of the proposed *p*GAN. The blue pipeline shows how the input photo ***B*** is generated to the initial sketch ***D***. The orange pipeline uses the initial sketch ***D*** as input, and introduces the illumination layer to remap the illumination on the face region ***E***. The green pipeline is used to refine the remapped sketch ***F*** and generates the final sketch ***G***. The objective function contains the adversarial loss and content loss between ***D***, ***G*** and ***H*** as detailed in Sec.3.4.

trast, artists subconsciously identify individual face components (*e.g.* hair, eyes, mouse, *etc.*) when drawing portraits. We argue that robust face sketch synthesis in the wild should satisfy:

- **To incorporate the information of facial components**. Thus, corresponding prior for individual components can be integrated. Such component priors are essential, *e.g.*, in removing background clutters.
- **To adjust illumination variations**. Causing via different lighting conditions and ethnic origins, illumination variation is regarded as the key issues in sketch synthesis. To that effect, one common operation is the contrast enhancement. For instance, the contrast limited adaptive histogram equalization (CLAHE) [Pizer *et al.*, 1987] is used to reduce the side lighting [Zhang *et al.*, 2010]. However, different scheme is only workable for its specific lighting condition, whose prior is hard to model beforehand.
- **To generate sketches with high speed and quality**. The traditional schemes [Zhou *et al.*, 2012; Wang *et al.*, 2017] fail in this goal, which resorts to either the time consuming operation (nearest neighbor selection) or the simple regression model.

In this paper, we propose a novel generative adversarial network termed *p*GAN, which efficiently generates face sketches against various imaging conditions using controlled training data. Figure 2 shows the workflow of the proposed *p*GAN model.

The key innovation lies in introducing important prior information into the process of synthesis and a designed parametric sigmoid activation function, which well deals with the aforementioned challenges. In particular, we design a stacked structure which connects a Resnet based generator [Zhu *et al.*, 2017a], a proposed illumination layer and an Unet generator [Isola *et al.*, 2016] together. A fully convolutional based discriminator [Ioffe and Szegedy, 2015] is used to achieve generative adversarial training. We embed background information and semantic components into the Resnet based generator and illumination layer, respectively. After remapping the illumination on the output of the Resnet based generator, we use the Unet generator to refine the remapped sketch and get the final sketch. Our contributions are summarized below:

- We propose a novel GAN based framework (termed *p*GAN), which can generate face sketches efficiently against imaging variations using controlled training data.
- We embed key photo priors into the process of synthesis and design a parametric sigmoid activation function, which can be regarded as a flexible contrast enhancement function to form a illumination layer.
- Perceptive and quantitative experiments demonstrate the outstanding performance and considerable generalization ability of the proposed *p*GAN.

2 Related Work

Despite the extensive progress in face sketch synthesis, most existing methods are still fail when dealing with the synthesis in the wild, *i.e.*, under uncontrolled conditions. Solving such synthesis problem has attracted ever-increasing research focus [Zhang *et al.*, 2010; Zhang *et al.*, 2015b; Peng *et al.*, 2016; Zhang *et al.*, 2017; Song *et al.*, 2017; Zhu *et al.*, 2017b]. We briefly review such works in this section.

To handle the lighting and pose variations, [Zhang *et al.*, 2010] proposed a robust face sketch synthesis algorithm termed MRF+. The invariance to lighting and pose conditions is achieved by adopting shape priors specific to facial components, accompanies by the usage of patch descriptors (*i.e.* SIFT). [Song *et al.*, 2017] proposed bidirectional illumination remapping (BLR), which was incorporated into data-driven synthesis methods to improve their robustness to lighting and pose variations. In particular, CLAHE, gamma correction and facial symmetry prior were used to reduce the effect of side lighting. Facial landmark and local affine transform were applied to handle pose variations. And [Zhang *et al.*, 2017] proposed an end-to-end photo-sketch mapping through structure and texture decomposition, which utilized nonparametric prior (*i.e.* the average of training sketches) and facial components to handle the lighting variations.

To synthesize photos with different backgrounds and non-facial factors (*e.g.* hairpins and glasses), [Zhang *et al.*, 2015b] presented a sparse representation-based greedy search (SRGS) scheme. Its principle is to search similar patch candidates globally by using patch descriptors with sparse representation and prior knowledge like image intensity and gradient. [Peng *et al.*, 2016] and [Zhu *et al.*, 2017b] modeled sketch synthesis as a multi-representation (MR) learning problem, which are robust to lighting variations and clutter backgrounds. In detail, [Peng *et al.*, 2016] extracted patch intensities, SURF and multi-scale LBP to obtain the multiple representations and [Zhu *et al.*, 2017b] extracted feature maps (deep representations) from a pre-trained 16-layer VGG net to represent the input photos.

In summary, aforementioned methods were designed to handle robust face sketch synthesis. However, the fast syn-

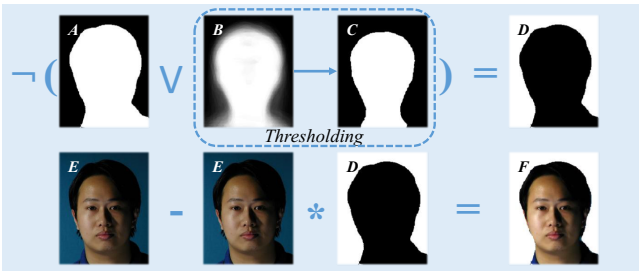

Figure 3: Illustration of background subtraction. The average mask ***B*** of training set is thresholded to a binary mask ***C*** to rectify the original mask ***A*** of image ***E***. The effect of image background is excluded by using the final binary mask ***D*** to obtain the final result ***F***.

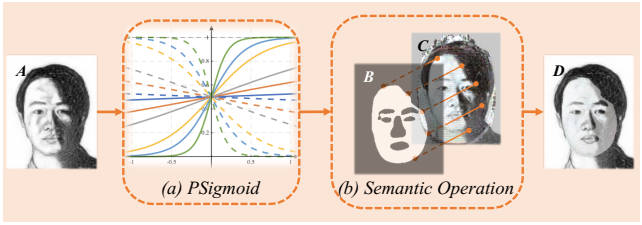

Figure 4: Illustration of illumination layer. The illumination layer consists of two parts: (a) PSigmoid, which is a contrast enhancement function to remap illumination on initial sketch ***A*** to obtain intermediate result ***C***; (b) Semantic operation. For example, we replace the face area ***B*** of initial sketch ***A*** with the corresponding illumination remapped area of intermediate result ***C***.

thesis methods of [Zhang *et al.*, 2015b] and [Zhang *et al.*, 2017] lack keeping the sketch style, and [Song *et al.*, 2017] is an inconvenient preprocessing for data-driven methods, while the other methods [Zhang *et al.*, 2010; Peng *et al.*, 2016; Zhu *et al.*, 2017b] generate sketches very slowly. In addition, these methods only deal with certain aspects, such as the lighting and pose variations or the background and non-facial factors. Thus, face sketch synthesis in the wild towards practical applications is still an open problem.

3 Robust Face Sketch Synthesis

3.1 Preliminaries

Given a face photo X with different imaging conditions against training photos, the trained model targets at efficiently generating a sketch portrait Y, which preserves the identity with a sketch style. To this end, we embed key photo priors into the process of synthesis and design a parametric sigmoid activation function for compensating illumination variations. The generative adversarial training is settled, where the Resnet based generator, the proposed illumination layer, the Unet generator and a fully convolutional based discriminator are used. After training, the generators and illumination layer are used for online sketch synthesis. In particular, we use prior information to assist the process of synthesis, as detailed in Sec.3.2. We further introduce our solution to handle the illumination variations in Sec.3.3 (*i.e.*, a learnable illumination layer), followed by the overall design of *p*GAN model in Sec.3.4.

3.2 Prior Information

Given an input photo X, we first introduce the prior information used as below. Specifically, we apply PortraitFCN proposed by [Shen *et al.*, 2016] to implement automatic portrait segmentation, which produces a score map that indicates the probability of a given pixel belonging to a subject. As shown in Figure 3, the score map is thresholded to a binary mask to subtract background. To reduce the segmentation error, we add the average mask of training set as a nonparametric prior. We also use the face parsing method (P-net) proposed by [Liu *et al.*, 2015] to decompose the input photo X into several semantic components (*e.g.* face, hair, eyes, mouse, *etc.*). We incorporate the binary mask of semantic component into the illumination layer to remap illumination on corresponding area. Figure 4(b) shows an example of semantic operation, in which we replace the original face area with the corresponding illumination remapped area. Inspired by the work of [Zhu *et al.*, 2017b], we regard the feature maps of the Resnet based generator as deep representations to improve the robustness against lighting variations.

3.3 Illumination Layer

Illumination variation is typically caused via different lighting conditions and ethnic origins. Widely used solution can resort to contrast enhancement, which can be defined as the slope of the function mapping input pixels to output pixels. To flexibly remap illumination is not a trivial task in convolutional net, since the slope of the standard activations like tanh, sigmoid or ReLU, is not adjustable. Towards learning an adjustable remapping function in neural network to model the flexible contrast enhancement, we propose a new extension of sigmoid termed Parametric Sigmoid (PSigmoid) as shown in Figure 4(a). This activation function adaptively learns the parameter of traditional sigmoid units, defined as:

$$f(x) = \frac{1}{1 + e^{-mx}}, \tag{1}$$

where x is the input of the nonlinear activation f, and m is a learnable parameter controlling the slope of the function. When $m = 1$, PSigmoid degenerates to the original sigmoid.

PSigmoid can be trained using backpropagation with other layers jointly. The updating of m is simply derived from the chain rule as follows:

$$\frac{\partial \mathcal{L}}{\partial m} = \frac{\partial \mathcal{L}}{\partial f(x)} \frac{\partial f(x)}{\partial m}, \tag{2}$$

where $\mathcal{L}$ denotes the objective function, the term $\frac{\partial \mathcal{L}}{\partial f(x)}$ is the gradient propagated from the deeper layers. The gradient of the PSigmoid is given by:

$$\frac{\partial f(x)}{\partial m} = x f(x)\big(1 - f(x)\big). \tag{3}$$

The PSigmoid can be combined with facial component masks. As shown in Figure 4(b), we take the face area as an example. Thus the illumination layer consists of two parts: PSigmoid, and Semantic operation shown in Figure 4(b).

3.4 The Proposed *p*GAN

As shown in Figure 2, the proposed *p*GAN consists four components, *i.e.*, (1) a Resnet based generator that extracts expressive deep representations, (2) an Unet generator that preserves the same underlying structure, (3) the illumination layer depicted in Sec.3.3 and (4) a fully convolutional based discriminator. Firstly, *p*GAN preprocesses an input photo by subtracting the background using the binary mask depicted in Sec.3.2. Secondly, a Resnet based generator is used to generate the initial sketch of the preprocessed photo. Thirdly, illumination remapping of the initial sketch is conducted as depicted in Sec.3.3. Finally, an Unet generator is employed to refine the sketch after the illumination remapping.

We have discovered that the illumination layer can do a better job when it works on image intensities instead of image features. So we force the Resnet based generator to generate an initial sketch instead of latent feature maps, which is then used as the input of the illumination layer. To this end, our loss function contains two terms to generate the initial sketch Y_R and the final sketch Y_U. Correspondingly, the adversarial losses $\mathcal{L}_{adv}$ for Y_R and Y_U are defined as follows:

$$\begin{aligned}\mathcal{L}_{adv}(R, D) &= \mathbb{E}_{x\sim p_{data}(x)}[logD(Y) + log(1 - D(R(X)))] \\ &= \mathbb{E}_{x\sim p_{data}(x)}[logD(Y) + log(1 - D(Y_R))], \quad (4)\end{aligned}$$

$$\begin{aligned}\mathcal{L}_{adv}(R, P, U, D) &= \mathbb{E}_{x\sim p_{data}(x)}[logD(Y) + log(1 - D(RPU(X)))] \\ &= \mathbb{E}_{x\sim p_{data}(x)}[logD(Y) + log(1 - D(Y_U))], \quad (5)\end{aligned}$$

where Y denotes the target sketch, R, P, U and D denote the Resnet based generator, the illumination layer, the Unet generator and the discriminator, respectively.

Since the normalized $\mathcal{L}_1$ distance introduces less blurring than the $\mathcal{L}_2$ distance, it is used to compute the content loss. The content losses $\mathcal{L}_c$ for Y_R and Y_U are defined as follows:

$$\mathcal{L}_c(R) = \|R(X) - Y\| = \|Y_R - Y\|, \quad (6)$$

$$\mathcal{L}_c(R, P, U) = \|RPU(X) - Y\| = \|Y_U - Y\|. \quad (7)$$

The overall objective function of *p*GAN is formulated as:

$$\begin{aligned}R^*, P^*, U^* = \arg \min_{R,P,U} \max_{D} \big(&\mathcal{L}_{adv}(R, D) + \lambda\mathcal{L}_c(R) \\ &+ \mathcal{L}_{adv}(R, P, U, D) + \lambda\mathcal{L}_c(R, P, U)\big), \quad (8)\end{aligned}$$

where λ is to balance the adversarial loss and the content loss.[1] We separate the overall objective function to optimize the discriminator, the illumination layer and the generators:

$$D^* = \arg \max_{D} \big(logD(Y) + log(1 - D(Y_R))\big), \quad (9)$$

$$R^* = \arg \min_{R} \big(log(1 - D(Y_R)) + \lambda\|Y_R - Y\|\big), \quad (10)$$

$$D^* = \arg \max_{D} \big(logD(Y) + log(1 - D(Y_U))\big), \quad (11)$$

$$\begin{aligned}R^*, P^*, U^* = \arg \min_{R,P,U} \big(&log(1 - D(Y_U)) \\ &+\lambda\|Y_U - Y\|\big). \quad (12)\end{aligned}$$

Algorithm 1 Optimization procedure of *p*GAN

Input: Initialized discriminator D, generator R, U and illumination layer P.
Output: Optimized D, R, U, P.

```
for i = 1 to nEpoches do
    Generate fake sketch Y_R by R.
    Optimize D by solving Eq.9.
    Fix D and optimize R by solving Eq.10.
    Generate fake sketch Y_U by R, P, U.
    Optimize D again by solving Eq.11.
    Fix D and optimize R, P, U by solving Eq.12.
end for
```

The optimization procedure of the proposed *p*GAN is shown in Algorithm 1.

4 Experiments

To quantize the performance of the proposed *p*GAN, we conduct experiments on the following datasets: the Chinese University of Hong Kong (CUHK) face sketch dataset (CUFS), the CUHK face sketch FERET (CUFSF) dataset [Phillips *et al.*, 2000], a set of Chinese celebrity photos obtained from the Web [Zhang *et al.*, 2010], a lighting variation set [Zhang *et al.*, 2010], and a pose variation set [Zhang *et al.*, 2010]. The CUFS dataset includes three subsets: the CUHK student dataset [Wang and Tang, 2009] (188 subjects), the AR dataset [Martinez, 1998] (123 subjects) and the XM2VTS dataset [Messer *et al.*, 1999] (295 subjects). Each subject has a face photo in a frontal pose under a normal lighting condition and a face sketch drawn by the artist. Photos in the CUFS dataset are different in ages, ethnic origins, genders, and backgrounds. Face photos in the CUFSF dataset are with illumination variation, and sketches are with shape exaggeration. The photos of Chinese celebrity have various backgrounds, lightings and poses. The photos in the lighting variation set have three different lightings (dark frontal/dark left/dark right) and the photos in the pose variation set include left and right poses with 45 degrees.

4.1 Cross-Dataset Experiments

To compare with existing methods, we select 88 subjects in the CUHK student dataset as the training set, while the rest 518 subjects are used as the testing set, which includes 123 photos from the AR dataset, 295 photos from the XM2VTS dataset and the rest 100 photos from the CUHK student dataset. It is noted that the CUHK student dataset, the AR dataset and the XM2VTS dataset are captured in different backgrounds, lightings and ethnic origins. Figure 5 shows comparisons of synthesis results with Multiple Representations-based method (MR) [Peng *et al.*, 2016], Markov random field (MRF) [Wang and Tang, 2009], Markov weight field (MWF) [Zhou *et al.*, 2012], spatial sketch denoising (SSD) [Song *et al.*, 2014], sparse representation-based greedy search (SRGS) [Zhang *et al.*, 2015b], branched fully convolutional network (BFCN) [Zhang *et al.*, 2017] and

[1] We empirically set λ as 100 in our implementation, which performs consistently better than other settings as we validated.

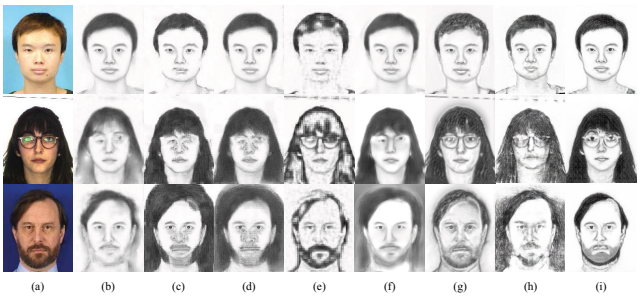

Figure 5: Comparison of sketches generated by different methods. (a) Input photo. (b) MR. (c) MRF. (d) MWF. (e) SRGS. (f) SSD. (g) BFCN. (h) cGAN. (i) *p*GAN. Photos listed in the first, second and third rows are from the CUHK student, AR and XM2VTS datasets, respectively.

Figure 6: Examples of synthesized sketches on the CUHK face sketch FERET database (CUFSF) by using the CUHK student dataset as the training set. The first row shows the test photos from the CUFSF dataset. The second row is the corresponding sketches drawn by the artist. The last row is the synthesized sketches of our method by training on the CUHK student dataset.

conditional GAN (cGAN) [Isola *et al.*, 2016]. As shown in Figure 5, the results from the other methods contain noise on the nose, hair and background, while the proposed *p*GAN performs much better on such aspects. In order to validate the generalization of our approach, we design the following experiments. As shown in Figure 6, the proposed method can deal with photos from the CUFSF dataset by using the CUHK student dataset as the training set. To further validate the generative ability of the proposed method across different training sets. Figure 7 shows the sketch synthesis results of the proposed method on various datasets by taking the AR and XM2VTS datasets as the training set, respectively.

4.2 Lighting and Pose Variations

We conduct experiments on the lighting variation set and the pose variation set. In our experiments, 88 persons from the CUHK student dataset are selected for training. Photos in the lighting variation set and the pose variation set are used for testing. Figure 8 shows some synthesized results from different methods, *i.e.*, MRF, MRF+ [Zhang *et al.*, 2010], SRGS, BFCN, cGAN and the proposed *p*GAN. Clearly, the results obtained by MRF, SRGS, BFCN and cGAN have noise, blur and artifacts. In contrast, MRF+ and *p*GAN can achieve satisfactory results. However, MRF+ is time-consuming, which limits its real-world application in efficient scenarios.

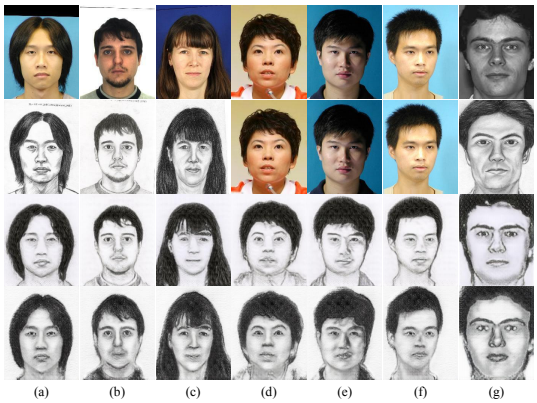

Figure 7: Examples of synthesized sketches on different datasets including (a)-(c) the CUHK database, (d) the Chinese celebrity photos, (e) the lighting variation set, (f) the pose variation set, and (g) the CUFSF database. The first row lists the test photos. The second row is the corresponding sketches drawn by the artist. The third row is the synthesized sketches of our method with the AR dataset as the training set. The last row is the synthesized sketches of our method with the XM2VTS dataset as the training set.

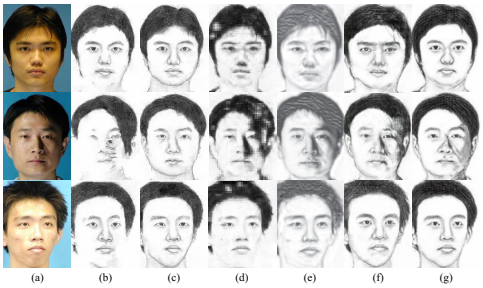

Figure 8: Comparison of the robustness to lighting and pose variations of different methods. (a) Input photo. (b) MRF. (c) MRF+. (d) SRGS. (e) BFCN. (f) cGAN. (g) *p*GAN. Photos in the first two rows come from the lighting variation set, and photo in the last row comes from the pose variation set.

4.3 Face Sketch Synthesis of Celebrity

The robustness of the proposed method is further demonstrated on a challenging set of face photos, which are collected from the Web with uncontrolled lightings, poses and background variations. In our experiments, 88 persons from the CUHK student dataset are selected for training. And photos in the set of Chinese celebrity photos are selected for testing. As shown in Figure 9, MRF, MWF, SRGS and cGAN usually produce distortions and noisy facial details, which is due to the large variations of skin color and the uncontrolled illumination. Although MRF+ and MR perform well compared to MRF, MWF, SRGS and cGAN, they still lose some sketch styles in hair regions compared to our method. And similar to the evaluation in Sec.4.2, MRF+ and MR are time-consuming while *p*GAN is quite fast.

4.4 Quantitative Evaluation

We use the feature similarity index (FSIM) [Zhang *et al.*, 2011] to further quantize the synthesis performance of dif-

Methods	LLE	MRF	MWF	SSD	Trans	SRGS	*p*GAN
CUHK	51.00(109)	59.47(108)	71.80(116)	68.40(103)	68.67(117)	76.93(84)	**82.93**(112)

Table 2: NLDA face recognition accuracy based on synthesized results from the CUHK dataset.

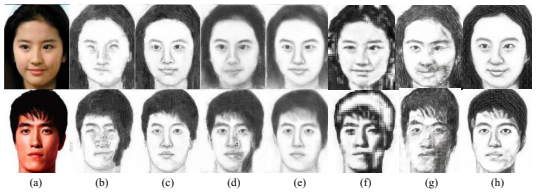

Figure 9: Synthesized results of celebrity photos from the Web. (a) Input photo. (b) MRF. (c) MRF+. (d) MWF. (e) MR. (f) SRGS. (g) cGAN. (h) *p*GAN.

Methods	MRF	MWF	SRGS	SSD	cGAN	*p*GAN
AR	68.94	72.48	68.99	71.30	68.21	**73.02**
XM2VTS	60.49	63.13	66.04	64.15	59.43	**67.32**

Table 1: FSIM values of different methods.

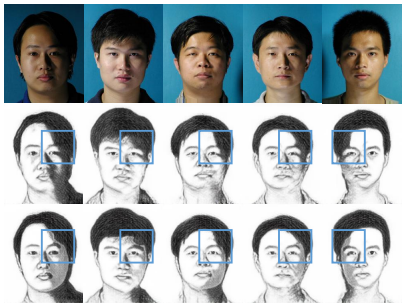

Figure 10: Comparison on models trained without/with the illumination layer. The first row shows the input photos under different lightings. The second row shows the synthesized results without the illumination layer. The third row shows the synthesized results with the illumination layer.

ferent methods, which is closely related to the subjective evaluation of human [Wang, 2014]. Specifically, we collect the synthesized results in Sec.4.1 on the AR dataset and the XM2VTS dataset. There are 418 (123 in the AR dataset and 295 in the XM2VTS dataset) synthesized sketches for each method. Table 1 gives the average FSIM scores on the AR dataset and the XM2VTS dataset, respectively. It can be seen that the proposed *p*GAN clearly outperforms five other methods. Face sketch recognition is frequently utilized to quantitatively evaluate the quality of the synthesized sketches. However, the accuracy of the recognition depends on various factors, such as the visual quality of the synthesized sketches, as well as the recognition models we apply. In this paper, we exploit the Null-space linear discriminant analysis (NLDA) as the recognition model to validate the proposed pGAN method. For the CUHK dataset, we obtain 418 synthesized sketches in total. We randomly choose 118 synthesized sketches and the corresponding original sketches as the training set to learn the classifier. The rest 300 synthesized sketches and the corresponding 300 original sketches are utilized as the test set. We repeat the recognition experiment 5 times by randomly dividing the 418 synthesized sketches into the training set and the test set. Table 2 shows the best recognition rate at a certain dimension. As shown in Table 2, our recognition accuracy is superior comparing to other methods.

4.5 Effectiveness of Illumination Layer

We conduct an experiment to verify the effectiveness of the illumination layer. Specifically, we remove the illumination layer to train the rest of the model, *i.e.*, without the orange pipeline shown in Figure 2. The results are depicted in the second row of Figure 10. For comparison, we depict the results obtained from the original model, whose results are shown in the third row of Figure 10. Obviously, the sketches generated with illumination layer are more robust to illumination variations.

4.6 Time Cost

For the traditional face sketch synthesis methods, the nearest neighbor selection (NNS) component is the most time-consuming part. Thus, we only discuss the computational complexity of the NNS involved in the existing methods. The time complexity of MRF, MRF+, MWF, SSD is $\mathrm{O}\left(cp^2MN\right)$ while that of SRGS, MR is $\mathrm{O}\left(C\right)$. For BFCN and *p*GAN, since there is no NNS part, the time complexity is $\mathrm{O}\left(1\right)$. Here c is the number of all possible candidates in the search region, p is the patch size, M is the number of image patches on each image, N is the number of training sketch-photo pairs, C is the number of local clusters.

5 Conclusion

We proposed a novel generative adversarial network termed *p*GAN for robust face sketch synthesis. The merit of *p*GAN is attributed to key photo priors embedded into the process of synthesis and a parametric sigmoid activation function designed for compensating illumination variations. The experiments demonstrate that *p*GAN can generate satisfactory results at low computational cost under uncontrolled conditions, such as different illumination, pose, background and ethnic origin. The proposed *p*GAN can handle face sketch synthesis in the wild towards practical applications.

Acknowledgements

This work is supported by the National Key R&D Program (No.2017YFC0113000, and No.2016YFB1001503), Nature Science Foundation of China (No.U1705262, No.61772443, and No.61572410), Post Doctoral Innovative Talent Support Program under Grant BX201600094, China Post-Doctoral Science Foundation under Grant 2017M612134, Scientific Research Project of National Language Committee of China

(Grant No. YB135-49), and Nature Science Foundation of Fujian Province, China (No. 2017J01125 and No. 2018J01106).

References

[Ioffe and Szegedy, 2015] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

[Isola *et al.*, 2016] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016.

[Liu *et al.*, 2005] Qingshan Liu, Xiaoou Tang, Hongliang Jin, Hanqing Lu, and Songde Ma. A nonlinear approach for face sketch synthesis and recognition. In *Computer Vision and Pattern Recognition*, pages 1005–1010, 2005.

[Liu *et al.*, 2015] Sifei Liu, Jimei Yang, Chang Huang, and Ming-Hsuan Yang. Multi-objective convolutional learning for face labeling. In *Computer Vision and Pattern Recognition*, pages 3451–3459, 2015.

[Martinez, 1998] Aleix M. Martinez. The ar face database. *CVC technical report*, 1998.

[Messer *et al.*, 1999] Kieron Messer, Jiri Matas, Josef Kittler, Juergen Luettin, and Gilbert Maitre. Xm2vtsdb: The extended m2vts database. In *Second international conference on audio and video-based biometric person authentication*, pages 965–966, 1999.

[Peng *et al.*, 2016] Chunlei Peng, Xinbo Gao, Nannan Wang, Dacheng Tao, Xuelong Li, and Jie Li. Multiple representations-based face sketch–photo synthesis. *IEEE transactions on neural networks and learning systems*, 27(11):2201–2215, 2016.

[Phillips *et al.*, 2000] P. Jonathon Phillips, Hyeonjoon Moon, Syed /A Rizvi, and Patrick J Rauss. The feret evaluation methodology for face recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000.

[Pizer *et al.*, 1987] Stephen M. Pizer, E. Philip Amburn, John D. Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B. Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*, 39(3):355–368, 1987.

[Shen *et al.*, 2016] Xiaoyong Shen, Aaron Hertzmann, Jiaya Jia, Sylvain Paris, Brian Price, Eli Shechtman, and Ian Sachs. Automatic portrait segmentation for image stylization. *Computer Graphics Forum*, 35(2):93–102, 2016.

[Song *et al.*, 2014] Yibing Song, Linchao Bao, Qingxiong Yang, and Ming-Hsuan Yang. Real-time exemplar-based face sketch synthesis. In *European Conference on Computer Vision*, pages 800–813, 2014.

[Song *et al.*, 2017] Yibing Song, Jiawei Zhang, Linchao Bao, and Qingxiong Yang. Fast preprocessing for robust face sketch synthesis. In *International Joint Conferences on Artificial Intelligence*, pages 4530–4536, 2017.

[Wang and Tang, 2009] Xiaogang Wang and Xiaoou Tang. Face photo-sketch synthesis and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):1955–1967, 2009.

[Wang *et al.*, 2013] Nannan Wang, Dacheng Tao, Xinbo Gao, Xuelong Li, and Jie Li. Transductive face sketch-photo synthesis. *IEEE transactions on neural networks and learning systems*, 24(9):1364–1376, 2013.

[Wang *et al.*, 2014] Nannan Wang, Dacheng Tao, Xinbo Gao, Xuelong Li, and Jie Li. A comprehensive survey to face hallucination. *International journal of computer vision*, 106(1):9–30, 2014.

[Wang *et al.*, 2017] Nannan Wang, Mingrui Zhu, Jie Li, Bin Song, and Zan Li. Data-driven vs. model-driven: Fast face sketch synthesis. *Neurocomputing*, pages 1–8, 2017.

[Wang, 2014] Nannan Wang. Heterogeneous facial image synthesis and its applications. *Xidian University*, 2014.

[Zhang *et al.*, 2010] Wei Zhang, Xiaogang Wang, and Xiaoou Tang. Lighting and pose robust face sketch synthesis. In *European Conference on Computer Vision*, pages 420–433, 2010.

[Zhang *et al.*, 2011] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on image processing*, 20(8):2378–2386, 2011.

[Zhang *et al.*, 2015a] Liliang Zhang, Liang Lin, Xian Wu, Shengyong Ding, and Lei Zhang. End-to-end photo-sketch generation via fully convolutional representation learning. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 627–634, 2015.

[Zhang *et al.*, 2015b] Shengchuan Zhang, Xinbo Gao, Nannan Wang, Jie Li, and Mingjin Zhang. Face sketch synthesis via sparse representation-based greedy search. *IEEE transactions on image processing*, 24(8):2466–2477, 2015.

[Zhang *et al.*, 2017] Dongyu Zhang, Liang Lin, Tianshui Chen, Xian Wu, Wenwei Tan, and Ebroul Izquierdo. Content-adaptive sketch portrait generation by decompositional representation learning. *IEEE transactions on image processing*, 26(11):328–339, 2017.

[Zhou *et al.*, 2012] Hao Zhou, Zhanghui Kuang, and Kwan-Yee K. Wong. Markov weight fields for face sketch synthesis. In *Computer Vision and Pattern Recognition*, pages 1091–1097, 2012.

[Zhu *et al.*, 2017a] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017.

[Zhu *et al.*, 2017b] Mingrui Zhu, Nannan Wang, Xinbo Gao, and Jie Li. Deep graphical feature learning for face sketch synthesis. In *International Joint Conferences on Artificial Intelligence*, pages 3574–3580, 2017.