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

Shengchuan Zhang\textsuperscript{1,2}, Rongrong Ji\textsuperscript{1,2*,} Jie Hu\textsuperscript{1,2}, Yue Gao\textsuperscript{3}, Chia-Wen Lin\textsuperscript{4}

\textsuperscript{1} Fujian Key Laboratory of Sensing and Computing for Smart City, Xiamen University
\textsuperscript{2} School of Information Science and Engineering, Xiamen University
\textsuperscript{3} School of Software, Tsinghua University
\textsuperscript{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 pGAN, which can generate face sketches efficiently using training data under fixed conditions and handle the aforementioned uncontrolled conditions. In pGAN, 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.

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-
contrast, 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 pGAN, which efficiently generates face sketches against various imaging conditions using controlled training data. Figure 2 shows the workflow of the proposed pGAN 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 pGAN), 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 pGAN.

## 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 decompositon, 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 nonfacial 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-
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 nonfacial 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 pGAN 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 areas. 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:

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

where \( m \) is the input of the nonlinear function \( f \), 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:

\[
    \frac{\partial L}{\partial m} = \frac{\partial f(x)}{\partial m} \frac{\partial f(x)}{\partial L},
\]

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

\[
    \frac{\partial f(x)}{\partial m} = x f(x) (1 - f(x)).
\]

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 pGAN

As shown in Figure 2, the proposed pGAN consists of 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, pGAN pre-processes 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 pre-processed 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 $L_{adv}$ for $Y_R$ and $Y_U$ are defined as follows:

$$L_{adv}(R, D) = \mathbb{E}_{x \sim p_{data}(x)} [\log D(Y) + \log(1 - D(R(X)))]$$

$$= \mathbb{E}_{x \sim p_{data}(x)} [\log D(Y) + \log(1 - D(Y_R))], \quad (4)$$

$$L_{adv}(R, P, U, D) = \mathbb{E}_{x \sim p_{data}(x)} [\log D(Y) + \log(1 - D(RPU(X)))]$$

$$= \mathbb{E}_{x \sim p_{data}(x)} [\log D(Y) + \log(1 - D(Y_U))], \quad (5)$$

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 $L_1$ distance introduces less blurring than the $L_2$ distance, it is used to compute the content loss. The content losses $L_c$ for $Y_R$ and $Y_U$ are defined as follows:

$$L_c(R) = \| R(X) - Y \| = \| Y_R - Y \|, \quad (6)$$

$$L_c(R, P, U) = \| RPU(X) - Y \| = \| Y_U - Y \|, \quad (7)$$

The overall objective function of pGAN is formulated as:

$$R^*, P^*, U^* = \arg \min_{R, P, U} \max_{D} (L_{adv}(R, D) + \lambda L_c(R)) + L_{adv}(R, P, U, D) + \lambda L_c(R, P, U), \quad (8)$$

where $\lambda$ is to balance the adversarial loss and the content loss. We separate the overall objective function to optimize the discriminator, the illumination layer and the generators:

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

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

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

$$R^*, P^*, U^* = \arg \min_{R, P, U} \log(1 - D(Y_U))$$

$$+ \lambda \| Y_U - Y \|. \quad (12)$$

Algorithm 1 Optimization procedure of pGAN

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


1: for $i = 1$ to $n\text{epochs}$ do
2: Generate fake sketch $Y_R$ by $R$.
3: Optimize $D$ by solving Eq.9.
4: Fix $D$ and optimize $R$ by solving Eq.10.
5: Generate fake sketch $Y_U$ by $R, P, U$.
6: Optimize $D$ again by solving Eq.11.
7: Fix $D$ and optimize $R, P, U$ by solving Eq.12.
8: end for

The optimization procedure of the proposed pGAN is shown in Algorithm 1.

4 Experiments

To quantize the performance of the proposed pGAN, 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 frontali/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 de-noising (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

Footnote:

\footnote{We empirically set $\lambda$ as 100 in our implementation, which performs consistently better than other settings as we validated.}
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 pGAN 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 database 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 pGAN. Clearly, the results obtained by MRF, SRGS, BFCN and cGAN have noise, blur and artifacts. In contrast, MRF+ and pGAN can achieve satisfactory results. However, MRF+ is time-consuming, which limits its real-world application in efficient scenarios.

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 pGAN 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-
different 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 pGAN 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 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 $O(cp^2MN)$ while that of SRGS, MR is $O(C)$. For BFCN and pGAN, since there is no NNS part, the time complexity is $O(1)$. 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 pGAN for robust face sketch synthesis. The merit of pGAN 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 pGAN can generate satisfactory results at low computational cost under uncontrolled conditions, such as different illumination, pose, background and ethnic origin. The proposed pGAN 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.
References


