High-Quality Synthesized Face Sketch Using Generative Reference Prior

Sami Mahfoud1,2, Messaoud Bengherabi2, Abdelhamid Daamouche3, Elhocine Boutella3 and Abdenour Hadid4

1 University of Algiers 3, Faculty of Economic, Commercial and Management Sciences, Laboratory of Governance and Modernization of Public Management, 02 Ahmed Ouaked Street Dely Ibrahim 16302, Algiers, Algeria.
2 Center for Development of Advanced Technologies, Telecom Division, P.O. Box 17 Baba-Hassen 16303, Algiers, Algeria.
3 University M’Hamed Bougara of Boumerdes, Institute of Electrical and Electronic Engineering, Laboratory of Signals and Systems, Boumerdes, 35000, Algeria.
4 Sorbonne University Abu Dhabi, Sorbonne Center for Artificial Intelligence, Abu Dhabi, UAE.

Abstract. Face sketch synthesis (FSS) is considered an image-to-image translation problem, where a face sketch is generated from an input face photo. FSS plays a vital role in video/image surveillance-based law enforcement. In this paper, motivated by the recent success of generative adversarial networks (GAN), we consider conditional GAN (cGAN) to approach the problem of face sketch synthesis. However, despite the powerful cGAN model’s ability to generate fine textures, low-quality inputs characterized by the facial sketches drawn by artists cannot offer realistic and faithful details and have unknown degradation due to the drawing process, while high-quality references are inaccessible or even nonexistent. In this context, we propose an approach based on Generative Reference Prior (GRP) to improve the synthesized face sketch perception. Our proposed model, that we call cGAN-GRP, leverages diverse and rich priors encapsulated in a pre-trained face GAN for generating high-quality facial sketch synthesis. Extensive experiments on publicly available face databases using facial sketch recognition rate and image quality assessment metrics as criteria demonstrate the effectiveness of our proposed model compared to several state-of-the-art methods.

Key words: Generative Adversarial Networks; Face Sketch Synthesis; Generative Reference Prior

1. INTRODUCTION

In security and law enforcement, police agencies can quickly identify potential suspects by automatically retrieving the suspect’s photos from the mugshot database [1]. Actually, the face photo of criminal suspects may not be available, and the face sketch of the likely suspect, drawn by expert artists based on the description of eyewitnesses, is an alternative way to assist in face sketch matching applications [2]. In addition to security applications, Face Sketch Synthesis (FSS) can be used in diverse digital entertainment applications. Face sketches are becoming more and more popular amongst social network users and smartphones, where face sketch is utilized as profile photos or avatars. However, face sketch synthesis and recognition may yield in challenging problems due to the significant discrepancy in texture and structure between the facial photo and facial sketch.

FSS techniques have been classified into two groups, namely data-driven and model-driven strategies [3]. Data-driven strategies synthesize facial sketch patches using a linear combination of similar training photo-sketch pairs. Model-driven strategies are generally based on the trained model, which can directly synthesize facial sketches from facial photos after learning an offline mapping function between two heterogeneous modalities.

Data-driven models synthesize a facial sketch of a test photo by linearly combining candidate facial sketch patches chosen from the training facial photo-sketch pairs. For instance, Tang and Wang [4] proposed synthesizing facial sketches using the principal component analysis technique to get the coefficients used in the synthesis process. However, the linear hypothesis restricted the capability to characterize the non-linear aspect between facial photos and sketches. Liu et al. [5] introduced a method via Locally Linear Embedding (LLE) to synthesize face sketches from face photos while the K-Nearest Neighbor (K-NN) is used to search similar neighbors at the image patch level. Zhang et al. [6] synthesized face sketches using similarity and prior knowledge between different facial image patches. Regarding the similarity between neighboring facial image patches, Wang and Tang [2] employed Markov Random Fields (MRF) in the FSS process and then utilized belief propagation to generate facial sketches. Zhou and al. [7] introduced a combination of K-neighbour patches into the MRF architecture named Markov Weight Fields (MWF), to resolve their deformation issue. Song et al. [8] proposed a fast synthesized method using Spatial Sketch Denoising (SSD) problem.

Deep learning, on the other hand, has recently been gaining interest and has been extensively applied to related problems such as image style transfer, image super-resolution, image classification, and image fusion [9, 10, 11, 12]. For the particular issue at hand, Zhang et al. [13] designed a model using a Convolutional Neural Network (CNN) to learn the end-to-end facial photo-to-facial sketch mapping. More recently, inspired by their significant contributions to different image-
2. PROPOSED METHOD

We describe below the overall architecture of our proposed cGAN-GRP approach along with a problem formulation, and details of each module of the proposed cGAN-GRP framework. Our proposed cGAN-GRP architecture combines two GAN models, namely the conditional GAN model for face sketch synthesis (cGAN) and Generative Reference Prior (GRP), to enhance the perceptual quality of the synthesized face sketch. The overall cGAN-GRP framework is depicted in Fig.2.

A. Conditional GAN model for the face sketch synthesis

We briefly describe the notations used to represent the cGAN for synthesized facial sketch. Depending on the \( M \) training facial sketch-photo pairs, the purpose is to produce the output \( s \), generated face sketch, from an observable facial image \( t \).

As shown in Fig.2 (cGAN model part), conditional Generative Adversarial Networks (cGAN) are a type of generative network that attempt to train a non-linear mapping function from the observable facial image \( t \) and a random noise vector \( z \) to create a sketch \( x \), cGAN : \( \{ t, z \} \rightarrow x \) instead of \( \{ z \} \rightarrow x \) as Generative Adversarial Networks (GANs) do.

The generator \( G \) attempts to produce fake face sketches that are unable to be differentiated by the discriminator \( D \) against the real sketches painted by professional artists. Simultaneously, the discriminator \( D \) attempts to differentiate the fake of the generator \( G \) among the real, as depicted in Fig.2. The conditional generative adversarial network objective [14] is expressed as follows:

\[
G^* = \arg \min_G \max_D \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{11}(G),
\]

where \( \mathcal{L}_{cGAN} \) is the cGAN loss, \( \mathcal{L}_{11}(G) \) is the regularization loss, and \( \lambda \) is utilized to make a balance within the \( \mathcal{L}_{cGAN} \) lose and the \( \mathcal{L}_{11}(G) \) loss. The \( \mathcal{L}_{cGAN} \) has the following definition:

\[
\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{x_{data} \sim P_{data}(x)}[\log D(x)] + \mathbb{E}_{z \sim P_{z}(z)}[\log (1 - D(t, G(z)))],
\]

and the \( \mathcal{L}_{11}(G) \) is expressed as follows:

\[
\mathcal{L}_{11}(G) = \mathbb{E}_{t \sim P_{data}(t), x \sim P_{x}(x)}[\| (x - G(1.z)) \|_1]
\]

The discriminator \( D \) in the cGAN model is exactly a classifier. Its aim is to learn to differentiate the real data from the fake data generated by \( G \). Any network appropriate to the facial image data could be used. For the discriminator, in our case, the PatchGAN [24] is utilized as the convolutional classifier. We adopt an architecture based on U-Net [25] as the generator \( G \).

B. Generative Reference Prior model

As shown in Fig. 1 and Fig.6, noise and blur attached to fine texture appear in the generated sketches by cGAN due to the pixel-to-pixel mapping on the one hand and the low-quality facial sketches drawn by artists and the unknown degradation caused by the drawing process on the other hand. To solve

![Fig. 1. Facial sketches synthesized by Cycle-GAN, cGAN and the proposed cGAN-GRP method. (Zoom in for the best view)](image-url)
this issue, we propose a new framework based on the Generative Reference Prior, which significantly reduces the noise and blurring in the synthesized face sketch. Our proposed GRP model takes advantage of extra and various prior information incorporated in a pre-trained facial StyleGAN2 to reconstruct realistic textures and structures, aiming to achieve a higher perceptual quality.

Given an input face photo (here, the synthesized facial sketch $x$ by the cGAN model), the Generative Reference Prior aims to assess a high-perceptual quality of the facial photo (here, the enhanced synthesized facial sketch $\hat{y}$) that resembles the high-quality ground truth photo $\hat{y}$ as closely as possible in terms of texture and structure, and realness.

Our proposed GRP is composed of two sub-models: the blur and noise removal module (such as U-Net) and the generative prior module (such as StyleGAN2). These two sub-modules are mapped by a code and several intermediate layers of Spatial Feature Transform (SFT). More details of these components are provided hereafter.

**Blur and noise removal module** is implemented as a U-Net architecture aiming to eliminate complicated degradation and obtain two features: multi-spatial $F_{Spat}$ features and the latent $F_{lat}$ features. The U-Net formulation is as follows:

$$F_{lat}, F_{Spat} = U-Net(x), \quad (4)$$

$F_{lat}$ maps the test photo into the latent code in the pre-trained StyleGAN2, whereas $F_{Spat}$ modulates StyleGAN2 features.

**Generative prior module** a pre-trained GAN model, such as StyleGAN2, encapsulates a learned distribution of facial features within its convolutional weights, known as the generative prior [27, 28]. We take advantage of these pre-trained facial GANs to generate diverse and rich facial details. One common method for exploiting generative priors involves mapping the input image to its most similar latent codes, denoted as $Z$, and subsequently producing the corresponding output using a pre-trained GAN [27, 28, 29]. Alternatively, intermediate convolutional features ($F_{GAN}$) of the closest face can be generated, as they offer more detail and can be adjusted by input features to enhance fidelity [22]. In our task, using the latter method, given the latent features $F_{lat}$ of the input photo (result of the U-Net, Eq. (4)), the first step is to map it to intermediate latent codes $W$, i.e., the intermediate space transformed from $Z$ with several Multi-Layer Perceptron (MLP) [29]. Subsequently, the latent codes $W$ generate GAN features for every resolution scale by passing through every convolution layer in the pre-trained StyleGAN2. the formulas are as follows:

$$W = MLP(F_{lat}),$$
$$F_{GAN} = StyleGAN(W). \quad (5)$$

Multi-spatial features $F_{Spat}$ are utilized to adjust the pre-trained facial GAN features $F_{GAN}$ to perform realistic results and faithful details while maintaining high fidelity. Specifically, we generate two transformation parameters ($\alpha, \beta$) from the input $F_{Spat}$ by several intermediate convolutional layers at each resolution scale. Hereafter, the vector $F_{GAN}$ is scaled and shifted to produce the modulation ($\alpha, \beta$). The formulas are expressed as follows:

$$\alpha, \beta = \text{Convolu}(F_{Spat}),$$
$$F_{output} = \text{SFT}(F_{GAN} | \alpha, \beta) = \alpha \odot F_{GAN} + \beta. \quad (6)$$

Consequently, SFT has the advantages of directly incorporating prior information and effective modulation of input images,
performing the best balance between texture faithfulness and fidelity to produce the final sketch $\hat{y}$.

C. Objective functions
The objective functions for training our GRP consist of i) reconstruction loss, ii) adversarial loss, and iii) identity preserving loss.

C.1. Reconstruction Loss: To restrict the output sketch $\hat{y}$ to be as similar as possible to the ground-truth facial photo $y$, we use the perceptual loss $\mathcal{L}_{\text{reco}}$ and the most utilized $L_1$ loss as the reconstruction loss function $\mathcal{L}_{\text{reco}}$, expressed below:

$$\mathcal{L}_{\text{reco}} = \lambda_{d_1} \|\hat{y} - y\|_1 + \lambda_{\text{perc}} \|\psi(\hat{y}) - \psi(y)\|_1,$$

where $\psi$ denotes the pre-trained VGG-19 model $\mathcal{L}_{\text{adve}}$ is the adversarial loss that represents the pre-trained VGG-19 model.

C.2. Adversarial Loss: We use this loss function $\mathcal{L}_{\text{adve}}$ to boost our GRP to generate realistic textures and structure. As in StyleGAN2 [23], we adopt the logistic loss $\mathcal{L}_{\text{adve}}$, expressed below:

$$\mathcal{L}_{\text{adve}} = -\lambda_{\text{adve}} \text{soft plus}(\mathcal{D}(\hat{y})), \tag{8}$$

where the loss weight is denoted by $\lambda$ and the discriminator is represented by $\mathcal{D}$.

C.3. Identity-Preserving Loss: We consider this loss function in the GRP architecture to encourage preserving identity. Similar to perceptual loss $\mathcal{L}_{\text{reco}}$ , we utilize the pre-trained ArcFace [33] network for its ability to capture the principal features required for discrimination. The identity-preserving loss $\mathcal{L}_{\text{ide}}$ is described as follows:

$$\mathcal{L}_{\text{ide}} = \lambda_{\text{ide}} \|v(\hat{y}) - v(y)\|_1, \tag{9}$$

where $v$ symbolises the pre-trained ArcFace [33] and $\lambda_{\text{ide}}$ indicates the loss weight.

C.4. Overall objective loss: The overall objective of GRP model is a sum of the aforementioned losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{reco}} + \mathcal{L}_{\text{adve}} + \mathcal{L}_{\text{ide}}. \tag{10}$$

The setting of loss weights are as follows: $\lambda_{d_1} = 0.1$, $\lambda_{\text{perc}} = 1$, $\lambda_{\text{adve}} = 0.1$ and $\lambda_{\text{ide}} = 10$.

Fig. 3. Some facial image examples and their sketches from the CUFS database. These three photo-sketch pairs are from the XM2VTS, CUHK Student, and AR datasets, respectively.
training. The remaining 188 synthesized facial sketches are used as the gallery test. By randomly partitioning the data, we repeat the process 20 times.

![Fig. 4. High-quality synthesized sketch and the improved hand-drawn face sketch using the proposed cGAN-GRP and GRP models, respectively. The photo and its corresponding hand-drawn sketch are from the CUFS database.](image)

A.4. Comparison with the state-of-the arts: For evaluating the effectiveness of the proposed strategy, we considered various state of the art methods for comparison, including LLE[5], SSD[8], MRF[2], MWF[7], Cycle-GAN[15], BP-GAN[16] and GAN[14]. The LLE, MWF, MRF, and SSD techniques are considered data-driven approaches, while the BP-GAN, cGAN, Cycle-GAN, and our proposed cGAN-GRP are model-driven.

B. Results and discussion

B.1. Quantitative evaluation: Fig.5 provides the statistics of LPIPS and Scoot values on the CUFS database as boxplots, and Table 1 shows the quantitative evaluation metrics for each approach used for comparison. The best results are highlighted in bold font. The results from Table 1 and Fig.5 demonstrate that our proposed cGAN-GRP attains the best values of LPIPS and Scoot while preserving the higher face sketch matching accuracy. Specifically, cGAN-GRP significantly increases the previous state-of-the-art Scoot by a considerable margin, 54.33%, achieved by cGAN method, to 63.39%. Besides, cGAN-GRP reduces the prior best state-of-the-art LPIPS from 22.49% to 21.57%. Such low LPIPS scores mean that the synthesized sketch by cGAN-GRP is more natural in terms of perceptual quality and appearance. In addition, the higher Scoot scores mean that the synthesized facial sketch by cGAN-GRP is more similar to those sketched by expert artists regarding textures and structures. In terms of face sketch matching, cGAN-GRP increases the accuracy from 95.53% to 95.66% as shown in Table 1.

According to these criteria, our cGAN-GRP demonstrates significant superiority over existing approaches and attains the best performance. Thus, our cGAN-GRP improves the perceptual quality of synthesized sketches while preserving high fidelity.

![Fig. 5. Scoot and LPIPS values of the different compared face sketch synthesis methods on the CUFS database, represented as box-plot curves, respectively.](image)

<table>
<thead>
<tr>
<th>FSS methods</th>
<th>Criteria</th>
<th>Scoot (%)</th>
<th>LPIPS(%)</th>
<th>NLDA(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data-driven methods</td>
<td>LLE</td>
<td>47.55</td>
<td>31.73</td>
<td>92.31</td>
</tr>
<tr>
<td></td>
<td>SSD</td>
<td>45.16</td>
<td>36.13</td>
<td>92.36</td>
</tr>
<tr>
<td></td>
<td>MRF</td>
<td>50.44</td>
<td>24.94</td>
<td>87.5</td>
</tr>
<tr>
<td></td>
<td>MWF</td>
<td>47.85</td>
<td>28.91</td>
<td>93.35</td>
</tr>
<tr>
<td>Model-driven (GAN-based) methods</td>
<td>BP-GAN</td>
<td>46.44</td>
<td>31.29</td>
<td>95.1</td>
</tr>
<tr>
<td></td>
<td>Cycle-GAN</td>
<td>50.04</td>
<td>30.07</td>
<td>86.06</td>
</tr>
<tr>
<td></td>
<td>cGAN</td>
<td>54.33</td>
<td>22.49</td>
<td>95.53</td>
</tr>
<tr>
<td></td>
<td>cGAN-GRP</td>
<td>63.39</td>
<td>21.57</td>
<td>95.66</td>
</tr>
</tbody>
</table>
B.2. Qualitative evaluation: Fig. 6 shows some synthesized facial sketches by the different techniques on the CUFS database. First, it can be illustrated from Fig. 6 that model-driven (BP-GAN, cGAN, Cycle-GAN, and our proposed cGAN-GRP) strategies generate finer details better than data-driven strategies, which include LLE, MWF, MRF, and SSD techniques. These outcomes explain the advantage of an adversarial process like the GAN model used for image-to-image translation problems. The data-driven approaches fail to provide fine details due to the existing nonlinear aspects between facial photos and facial sketches penciled by professional artists, including shape exaggeration, lighting variations, and different races. This justifies the blur and noise appearing on the generated face sketches using data-driven methods. Second, although the GAN-based methods, such as BP-GAN, cGAN, and Cycle-GAN, produce fine texture details, there is still some noise present among the facial parts, and the synthesized outcomes do not exhibit a net appearance, as depicted in Fig. 6. In contrast, the synthesized facial sketches using the proposed cGAN-GRP framework demonstrate a more realistic appearance and faithful details, as well as less noise. This is particularly more perceptible around the mouth, hair, and eye regions. Such superiority confirms the effectiveness of our proposed cGAN-GRP technique.

4. CONCLUSION

In this work, we adopted the conditional GAN for facial sketch synthesis. Then, we proposed enhancing the perceptual quality of synthesized facial sketches using the Generative Reference Prior strategy. Extensive experiment showed that despite the GAN-based approaches effectively maintaining texture and structure details, they also generate some noise and blur around the facial part, and some generated outputs are unclear. Our proposed cGAN-GRP model significantly reduces the noise, enhances the quality of the facial sketches, and provides more faithful and realistic details. Both quantitative and qualitative evaluations demonstrated that our proposed cGAN-GRP outperforms the current FSS techniques. In the proposed framework, the GRF and cGAN models have been separated. In the future, we plan to integrate the GRP into the cGAN architecture and generalize the proposed framework to other heterogeneous face recognition applications.

DECLARATION OF COMPETING INTERESTS

The authors certify that they have no known financial or interpersonal conflicts that could have been expected to have an impact on the research presented in this study.

ACKNOWLEDGEMENTS

This work is supported by the DGRSDT (General Directorate of Scientific Research and Technological Development), Algeria, the CDTA (Center for Development of Advanced Technologies), Algeria, and the GMPM (Laboratory of Governance and Modernization of Public Management), University of Algiers 3, Algeria.
REFERENCES


