Towards Realistic Face Photo-Sketch Synthesis via Composition-Aided GANs

Jun Yu, Senior Member, IEEE, Shengjie Shi, Fei Gao†, Dacheng Tao, Fellow, IEEE, and Qingming Huang, Fellow, IEEE

Abstract—Face photo-sketch synthesis aims at generating a facial sketch/photo conditioned on a given photo/sketch. It is of wide applications including digital entertainment and law enforcement. Precisely depicting face photos/sketches remains challenging due to the restrictions on structural realism and textural consistency. While existing methods achieve compelling results, they mostly yield blurred effects and great deformation over various facial components, leading to the unrealistic feeling of synthesized images. To tackle this challenge, in this work, we propose to use the facial composition information to help the synthesis of face sketch/photo. Specially, we propose a novel composition-aided generative adversarial network (CA-GAN) for face photo-sketch synthesis. In CA-GAN, we utilize paired inputs including a face photo/sketch and the corresponding pixel-wise face labels for generating a sketch/photo. In addition, to focus training on hard-generated components and delicate facial structures, we propose a compositional reconstruction loss. Finally, we use stacked CA-GANs (SCA-GAN) to further rectify defects and add compelling details. Experimental results show that our method is capable of generating both visually comfortable and identity-preserving face sketches/photos over a wide range of challenging data. Our method achieves the state-of-the-art quality, reducing best previous Fréchet Inception distance (FID) by a large margin. Besides, we demonstrate that the proposed method is of considerable generalization ability. We have made our code and results publicly available: https://fei-hdu.github.io/ca-gan/.

Index Terms—Face photo-sketch synthesis, image-to-image translation, generative adversarial network, deep learning.

I. INTRODUCTION

Face photo-sketch synthesis refers synthesizing a face sketch (or photo) given one input face photo (or sketch). It has a wide range of applications such as digital entertainment and law enforcement. Ideally, the synthesized photo or sketch portrait should be identity-preserved and appearance-realistic, so that it will yield both high sketch identification accuracy and excellent perceptual quality.

So far, tremendous efforts have been made to develop facial sketch synthesis methods, both shallow-learning based and deep-learning based [1], [2]. Especially, inspired by the great success of Generative Adversarial Networks (GANs) in various image-to-image translation tasks [3], researchers recently extended GANs for face sketch synthesis [4], [5], [6]. While these methods achieve compelling results, precisely depicting face sketches remains challenging due to the restrictions on structural realism and textural consistency. By carefully examining the synthesized sketches from existing methods, we observe serious deformations and aliasing defects over the mouse and hair regions. Besides, the synthesized photos/sketches are typically unpleasantly blurred. Such artifacts lead to the unrealistic feeling of synthesized sketches.

To tackle this challenge, we propose to use the facial composition information to help face photo-sketch synthesis. Specially, we propose to use pixel-wise face labelling masks to characterize the facial composition. This is motivated by the following two observations. First, pixel-wise face labelling masks are capable of representing the strong geometric constrain and complicated structural details of faces. Second, it is easy to access pixel-wise facial labels due to recent development on face parsing techniques [7], thus avoiding heavy human annotations and feasible for practical applications.

Additionally, we propose a composition-adaptive reconstruction loss to focus training on hard-generated components and prevents the large components from overwhelming the generator during training [8]. In typical image generation methods, the reconstruction loss is uniformly calculated across the whole image as (part of) the objective [3]. Thus large components that comprise a vast number of pixels dominate the training procedure, obstructing the model to generate delicate facial structures. However, for face photos/sketches, large components are typically unimportant for recognition (e.g. background) or easy to generate (e.g. facial skin). In contrast, small components (e.g. eyes) typically comprise complicated structures, and thus difficult to generate. To eliminate this barrier, we introduce a weighting factor for the distinct pixel loss of each component, which down-weights the loss assigned to large components.

We refer to the resulted model as Composition-Aided Generative Adversarial Network (CA-GAN). In CA-GAN, we utilize paired inputs including a face photo/sketch and the corresponding pixel-wise face labelling masks for generating the portrait. Besides, we use the improved reconstruction loss for training. Moreover, we use stacked CA-GANs (SCA-GAN) for refinement, which proves to be capable of rectifying
defects and adding compelling details [9]. As the proposed framework jointly exploits the image appearance space and structural composition space, it is capable of generating natural face photos and sketches. Experimental results show that our methods significantly outperform existing methods in terms of perceptual quality, and obtain better or comparable face recognition accuracies. We also verify the excellent generalization ability of our new model on faces in the wild.

The contributions of this paper are mainly three-fold.

- First, to the best of our knowledge, this is the first work to employ facial composition information in the loop of learning a face photo-sketch synthesis model.
- Second, we propose an improved pixel-wise reconstruction loss to focus training on hard-generated components and delicate facial structures, which is demonstrated to be much effective. Besides, it both speeds the training up and greatly stabilizes it.
- Third, our method achieves the state-of-the-art quality, reducing best previous Fréchet Inception distance (FID) by a large margin.

The rest of this paper is organized as follows. Section II introduces related works. Section III details the proposed method. Experimental results and analysis are presented in section IV. Section V concludes this paper.

II. RELATED WORK

A. Face Photo-Sketch Synthesis

Tremendous efforts have been made to develop facial photo-sketch synthesis methods, which can be broadly classified into two groups: data-driven methods and model-driven methods [10]. The former refers to methods that try to synthesize a photo/sketch by using a linear combination of similar training photo/sketch patches [11], [12], [13], [14], [15], [16]. These methods have two main parts: similar photo/sketch patch searching and linear combination weight computation. The similar photo/sketch searching process heavily increases the time consuming for test. Model-driven refers to methods that learn a mathematical function offline to map a photo to a sketch or inversely [17], [18], [19], [20]. Traditionally, researchers pay great efforts to explore hand-crafted features, neighbour searching strategies, and learning techniques. However, these methods typically yield serious blurred effects and great deformation in synthesized face photos and sketches.

Recently, a number of trials are made to learn deep learning based face sketch synthesis models. For example, Zhang et al. [21] propose to use branched fully convolutional network (BFCN) for generating structural and textural representations, respectively, and then use face parsing results to fuse them together. However, the resulted sketches exists heavily blurred and ring effects. More recently, inspired by the great success achieved by conditional Generative Network (cGAN) [22], [23] in various image-to-image translation tasks [24]. To name a few, Wang et al. [4] propose to first generate a sketch using the vanilla cGANs [3] and then refine it by using a post-processing approach termed back projection. Experimental results show that Pix2Pix can produces sketch-like structures in the synthesized portrait. Di et al. combine the Convolutional Variational Autoncoder and cGANs for attribute-aware face sketch synthesis [5]. However, there are also great deformation in various facial parts. Wang et al. follow the ideas of Pix2Pix [3] and CycleGAN [25], and use multi-scale discriminators for generating high-quality photos/sketches [26]. Most recently, Zhang et al. embed photo priors into cGANs and design a parametric sigmoid activation function for compensating illumination variations [6].

Few exiting methods use the composition information to guide the generation of the face sketch [21], [27] in a heuristic manner. In particular, they try to learn a specific generator for each component and then combine them together to form the entire face. Similar ideas have also been proposed for face image hallucination [28], [29]. In contrast, we propose to employ facial composition information in the loop of learning to boost the performance.

B. Image-to-image Translation

Our work is highly related to image-to-image translation, which has achieved significant progress with the development of generative adversarial networks (GANs) [23], [30] and variational auto-encoders (VAEs) [31]. Among them, conditional generative adversarial networks (cGAN) [3] attracts growing attentions because there are many interesting works based on it, including conditional face generation [32], text to image synthesis [9], and image style transfer [33]. Stacked networks have achieved great success in various directions, such as image super-resolution reconstruction [9] and image generation [34]. All of them obtained amazing results. Inspired by these observations, we are interested in generating sketch-realistic portraits by using cGAN. However, we found the vanilla cGAN [3] insufficient for this task, thus propose to boost the performance by both developing the network architecture and modifying the objective.

III. METHOD

A. Preliminaries

The proposed method is capable of handling both sketch synthesis and photo synthesis, because these two procedures are symmetric. In this section, we take face sketch synthesis as an example to introduce our method.

Our problem is defined as follows. Given a face photo $X$, we would like to generate a sketch portrait $Y$ that shares the same identity with sketch-realistic appearance. Our key idea is using the face composition information to help the generation of sketch portrait. The first step is to obtain the structural composition of a face. Face parsing can assign a compositional label for each pixel in a facial image. We thus employ the face parsing result (i.e. pixel-wise labelling masks) $M$ as prior knowledge for the facial composition. The remaining problem is to generate the sketch portrait based on the face photo and composition masks: $\{X, M\} \rightarrow Y$. Here, we propose a composition-aided GAN (CA-GAN) for this purpose. We further employ stacked CA-GANs (SCA-GAN) to refine the generated sketch portraits. Details will be introduced next.
B. Face Decomposition

Assume that the given face photo is \( X \in \mathbb{R}^{m \times n \times d} \), where \( m \), \( n \), and \( d \) are the height, width, and number of channels, respectively. We decompose the input photo into \( C \) components (e.g., hair, nose, etc.) by employing the face parsing method proposed by Liu et al. [7] due to its excellent performance. For notational convenience, we refer to this model as P-Net. By using P-Net, we get the pixel-wise labels related to 8 components, i.e., two eyes, two eyebrows, nose, upper and lower lips, inner mouth, facial skin, hair, and background [7].

Let \( \mathcal{M} = \{ M^{(1)}, \cdots, M^{(C)} \} \in \mathbb{R}^{m \times n \times C} \) denote the pixel-wise face labelling masks. Here, \( M^{(c)}_{i,j} \in [0,1] \) s.t. \( \sum_{c} M^{(c)}_{i,j} = 1 \) denotes that pixel \( X_{i,j} \) belongs to the \( c \)-th component, predicted by P-Net, \( c = 1, \cdots, C \) with \( C = 8 \). We use soft labels (probabilistic outputs) in this paper. In the preliminary implementation, we also tested our model while using hard labels (binary outputs), i.e., each value \( M^{(c)}_{i,j} \) denotes whether \( X_{i,j} \) belongs to the \( c \)-th component. Because it is almost impossible to get absolutely precise pixel-wise face labels, using hard labels occasionally yields deformation in the border area between adjacent components.

We note that an existing face parsing model [7] is adopted here, as this paper is mainly to explore how to use facial composition information to boost the performance of photo-sketch synthesis. We expect that an advanced face parsing model will be complementary to our approach, but it is beyond the scope of this paper.

C. Composition-aided GAN (CA-GAN)

In the proposed framework, we first utilize paired inputs including a face photo and the corresponding pixel-wise face labels for generating the portrait. Second, we propose a compositional reconstruction loss, to focus training on hard-generated components and delicate facial structures. Moreover, we use stacked CA-GANs to further rectify defects and add compelling details.

1) Generator Architecture: The architecture of the generator in CA-GAN is presented in Fig. 1. In our case, the generator needs to translate two inputs (i.e., the face photo \( X \) and the face labelling masks \( \mathcal{M} \)) into a single output \( \hat{Y} \). Because \( X \) and \( \mathcal{M} \) are of different modalities, we propose to use distinct encoders to model them and refer to them as Appearance Encoder and Composition Encoder, correspondingly. The outputs of these two encoders are concatenated at the bottleneck layer for the decoder [35]. In this way, the information of both facial details and composition can be well modelled respectively.

The architectures of the encoder, decoder, and discriminator are exactly the same as those used in [3] but without dropout, following the shape of a ”U-Net”. Specifically, we concatenate all channels at layer \( i \) in both encoders with those at layer \( n - i \) in the decoder. Details of the network can be found in the appendix of [3]. We note that we use the network in [3] here because it is a milestone in the image-to-image translation community and has shown appealing results in various tasks. Nevertheless, our proposed techniques are complementary for arbitrary cGANs frameworks.

In the preliminary experiment, we test the network with one single encoder that takes the cascade of \( X \) and \( \mathcal{M} \), i.e., \( [X, M^{(1)}, \cdots, M^{(C)}] \in \mathbb{R}^{m \times n \times (d + C)} \), as the input. This network is the most straightforward solution for simultaneously encoding the face photo and the composition masks. Experimental results show that using this structure decreases the face sketch recognition accuracy by about 2 percent and yield slightly blurred effects in the area of hair.

2) Compositional Reconstruction Loss: Previous approaches of cGANs have found it beneficial to mix the GAN objective with pixel-wise reconstruction loss for various tasks, e.g., image translation [3] and super-resolution reconstruction [24]. Besides, using the normalized \( L_1 \) distance encourage less blurring than the \( L_2 \) distance. We therefore use the normalized \( L_1 \) distance between the generated sketch \( \hat{Y} \) and the target \( Y \) in the computation of reconstruction loss. We introduce the compositional reconstruction loss starting from the standard reconstruction loss for image generation.

\[ L_{L_1,\text{global}}(Y, \hat{Y}) = \frac{1}{mn} \| Y - \hat{Y} \|_1, \]

In the global pixel loss, the \( L_1 \) loss related to the \( c \)-th component, \( c = 1, 2, \cdots, C \), can be expressed as:

\[ L_{L_1,\text{global}}^{(c)} = \frac{1}{mn} \| Y \odot M^{(c)} - \hat{Y} \odot M^{(c)} \|_1, \]

with \( L_{L_1,\text{global}} = \sum_{c} L_{L_1,\text{global}}^{(c)} \). Here, \( \odot \) denotes the pixel-wise product operation. As all the pixels are treated equally in the global reconstruction loss, large components (e.g. background and facial skin) contribute more to learn the generator than small components (e.g. eyes and mouth).

Compositional Loss. In this paper, we introduce a weighting factor, \( \gamma_c \), to balance the distinct reconstruction loss of each component. Specially, inspired by the idea of balanced cross-entropy loss [8], we set \( \gamma_c \) by inverse component frequency. When we adopt the soft facial labels, \( M^{(c)} \odot 1 \) is the sum of the
possibilities every pixel belonging to the $c^{th}$ component. Here, $\otimes$ denotes the convolutional operation. If we adopt the hard
facial labels, it becomes the number of pixels belonging to the
$c^{th}$ component. The component frequency is thus $\frac{M^{(c)}}{\sum_{c=1}^{C} M^{(c)}}$. So
we set $\gamma_{c} = \frac{\min_{c} M^{(c)}}{\sum_{c=1}^{C} M^{(c)}}$ and multiply it with $L_{L_{1,global}}^{(c)}$, resulting
in the balanced $L_{1}$ loss:

$$L_{L_{1,comp}}^{(c)} = \frac{1}{M^{(c)}} \sum_{c=1}^{C} \frac{L_{L_{1,comp}}^{(c)}}{1}\|Y \otimes M^{(c)} - \hat{Y} \otimes M^{(c)}\|_1$$

Obviously, the balanced $L_{1}$ loss is exactly the normalized $L_{1}$ loss across the related compositional region.

The compositional reconstruction loss is defined as,

$$L_{L_{1,comp}}(Y, \hat{Y}) = \sum_{c=1}^{C} L_{L_{1,comp}}^{(c)}$$

As $\gamma_{c}$ is broadly in inverse proportion to the component size, it reduces the loss contribution from large components. From the other aspect, it high-weights the losses assigned to small and hard-generated components. Thus the compositional loss focus training on hard components with tiny details, and prevents the vast number of pixels of unimportant component (e.g. background) or easy component (e.g. facial skin) from overwhelming the generator during training.

In practice, we use a weighted average of the global reconstruction loss and compositional reconstruction loss:

$$L_{L_{1}}(Y, \hat{Y}) = \alpha L_{L_{1,comp}} + (1 - \alpha) L_{L_{1,global}},$$

where $\alpha \in [0, 1]$ is used to balance the global reconstruction loss and the compositional pixel loss. We adopt this form in our experiments and set $\alpha = 0.7$, as it yields slightly improved perceptual comfortability over the compositional loss. In the following, we refer to the weighted reconstruction loss as compositional loss for facility.

3) Objective: We express the adversarial loss of CA-GAN as [3]:

$$L_{adv}(G, D) = E_{X, M, Y \sim p_{data}(X, M, Y)}[\log D(X, M, Y)] + E_{X, M \sim p_{data}(X, M)}[\log(1 - D(X, M, G(X, M)))]$$

Similar to the settings in [3], we do not add a Gaussian noise $z$ as the input.

Finally, we use a combination of the adversarial loss and the compositional loss to learn the generator. We aim to solve:

$$G^* = \arg\min_{G} \max_{D} L_{adv} + \lambda L_{L_{1}},$$

where $\lambda$ is a weighting factor.

D. Stacked Refinement Network

Finally, we use stacked CA-GAN (SCA-GAN) to further boost the quality of the generated sketch portrait [9]. The architecture of SCA-GAN is illustrated in Fig. 2.

SCA-GAN includes two-stage GANs, each comprises a generator and a discriminator, which are sequentially denoted by $G^{(1)}, D^{(1)}, G^{(2)}, D^{(2)}$. Stage-I GAN yields an initial portrait, $\hat{Y}^{(1)}$, based on the given face photo $X$ and pixel-wise label masks $M$. Afterwards, Stage-II GAN takes $\{X, M, \hat{Y}^{(1)}\}$ as inputs to rectify defects and add compelling details, yielding a refined sketch portrait, $\hat{Y}^{(2)}$.

The network architectures of the these two GANs are almost the same, except that the inputs of $G^{(2)}$ and $D^{(2)}$ have one more channel (i.e. the initial sketch) than those of $G^{(1)}$ and $D^{(1)}$, correspondingly. Here, the given photo and the initial sketch are concatenated and input into the appearance encoder.

E. Optimization

To optimize our networks, following [3], we alternate between one gradient descent step on $D$, then one step on $G$. We use minibatch SGD and apply the Adam solver. For clarity, we illustrate the optimization procedure of SCA-GAN in Algorithm 1. In our experiments, we use batch size 1 and run for 700 epochs in all the experiments. Besides, we apply instance normalization, which has shown great superiority over batch normalization in the task of image generation [3]. We trained our models on a single Pascal Titan Xp GPU. When we used a training set of 500 samples, it took about 3 hours to train the CA-GAN model and 6 hours to train the SCA-GAN model.

Algorithm 1 Optimization procedure of SCA-GAN (for sketch synthesis).

Input: a set of training instances, in form of triplet: 
\{a face photo $X$, pix-wise label masks $M$, a target sketch $Y$ \}; iteration time $t = 0$, max iteration $T$.

Output: optimal $G^{(1)}, D^{(1)}, G^{(2)}, D^{(2)}$; initial $G^{(1)}, D^{(1)}, G^{(2)}, D^{(2)}$;

for $t = 1$ to $T$ do
1. Randomly select one training instance:
\{a face photo $X$, pix-wise label masks $M$, a target sketch $Y$\}.
2. Estimate the initial sketch portrait:
\(Y^{(1)} = G^{(1)}(X, M)\).
3. Estimate the refined sketch portrait:
\(\hat{Y}^{(2)} = G^{(2)}(X, M, \hat{Y}^{(1)})\).
4. Update $D^{(1)}$:
\(D^{(1)+} = \arg\min_{D^{(1)}} L_{adv}(G^{(1)}, D^{(1)})\).
5. Update $D^{(2)}$:
\(D^{(2)+} = \arg\min_{D^{(2)}} L_{adv}(G^{(2)}, D^{(2)})\).
6. Update $G^{(1)}$:
\(G^{(1)+} = \arg\max_{G^{(1)}} L_{adv}(G^{(1)}, D^{(1)}) + \lambda L_{L_{1}}(Y, \hat{Y}^{(1)})\).
7. Update $G^{(2)}$:
\(G^{(2)+} = \arg\max_{G^{(2)}} L_{adv}(G^{(2)}, D^{(2)}) + \lambda L_{L_{1}}(Y, \hat{Y}^{(2)})\).

end for

IV. EXPERIMENTS

In this section, we will first introduce the experimental settings and then present a series of empirical results to verify the effectiveness of the proposed method.
A. Settings

1) Datasets: We conducted experiments on two widely used and public available datasets: the CUHK Face Sketch (CUFS) dataset [36] and the CUF SF dataset [37]. The composition of these datasets are briefly introduced below.

- The CUFS dataset consists of 606 face photos from three datasets: the CUHK student dataset [38] (188 persons), the AR dataset [39] (123 persons), and the XM2VTS dataset [40] (295 persons). For each person, there are one face photo and one face sketch drawn by the artist.
- The CUF SF dataset includes 1194 persons [41]. In the CUF SF dataset, there are lighting variation in face photos and shape exaggeration in sketches. Thus the CUF SF is very challenging. For each person, there are one face photo and one face sketch drawn by the artist.

Dataset partition. There are great divergences in the experimental settings among existing works. In this paper, we follow the most used settings presented in [1], and split the dataset in the following ways. For the CUFS dataset, 268 face photo-sketch pairs (including 88 pairs from the CUHK student dataset, 80 pairs from the AR dataset, and 100 pairs from the XM2VTS dataset) are selected for training, and the rest are for testing. For the CUF SF dataset, 250 pairs are selected for training, and the rest 944 pairs are for testing.

Preprocessing. Following existing methods [1], all these face images (photos and sketches) are geometrically aligned relying on three points: two eye centers and the mouth center. The aligned images are cropped to the size of 250 × 200. In the proposed method, the input image should be of fixed size, e.g., 256 × 256. In the default setting of [3], the input image is resized from an arbitrary size to 256 × 256. However, we observed that resizing the input face photo will yield serious blurred effects and great deformation in the generated sketch [4] [26]. In contrast, by padding the input image to the target size, we can obtain considerable performance improvement. We therefore use zero-padding across all the experiments.

2) Criteria: In this work, we choose the Fréchet Inception distance (FID) to evaluate the realism and variation of synthesized photos and sketches [42], [43]. FID has been widely used in image generation tasks and shown highly consistency with human perception. Lower FID values mean closer distances between synthetic and real data distributions. In our experiments, we use all the test samples to compute the FID, where the 2048-dimensional feature of the Inception-v3 network pre-trained on ImageNet is used [44].

We additionally adopt the Feature Similarity Index Metric (FSIM) [45] between a synthesized image and the corresponding ground-truth image to objectively assess the quality of the synthesized image. Notably, although FSIM works well for evaluating quality of natural images and has become a prevalent metric in the face photo-sketch synthesis community, it is of low consistency with human perception for synthesized face photos and sketches [46].

Finally, we statistically evaluate the face recognition accuracy while using the ground-truth photos/sketches as the probe image and synthesized photos/sketches as the images in the gallery. Null-space Linear Discriminant Analysis (NLDA) [47] is employed to conduct the face recognition experiments. We repeat each face recognition experiment 20 times by randomly splitting the data and report the average accuracy.

We use the proposed architecture for both sketch synthesis and photo synthesis. In the following context, we present a series of experiments:

- First, we perform ablation study on face photo-sketch synthesis on the CUFS dataset (see Part IV-B);
- Second, we perform face photo-sketch synthesis on the CUF SF and CUF SF datasets and compare with existing advanced methods (see Part IV-C and Part IV-D);
- Third, we conduct experiments on faces in the wild to verify whether the proposed method is robust to lighting and pose variations (see Part IV-E); and
- Finally, we verify that the proposed techniques stabilize the training procedure (see Part IV-F).

Our code and results are publicly available at: https://github.com/fei-hdu/ca-gan.

B. Ablation Study

We first evaluate the effectiveness of our design choices, including (i) using face composition (pixel-wise labels) as auxiliary input, (ii) the compositional loss, and (iii) stacked refinement. To this end, we construct several model variants and separately conduct photo synthesis and sketch synthesis experiments on the CUFS dataset. Results are shown in Table I and discussed next.

Face composition masks: Table I shows that using compositional masks as auxiliary input significantly improve the realism of the synthesized face images. Specially, it decreases FID by 2.7 (43.2 → 40.5) for sketch synthesis and by 36.5 (117.6 → 81.1) for photo synthesis. Besides, this improves the face recognition accuracy from 89.0 to 98.8 for photo synthesis, suggesting that compositional information is essential for photo-based face recognition.

Compositional loss: Table I shows that using compositional loss decreases FID by 3.0 (43.2 → 40.2) for sketch synthesis and by 34.5 (117.6 → 83.1) for photo synthesis. This highlights our motivation that focusing training on hard components is key.

CA-GAN uses both composition masks and compositional reconstruction loss. We note that CA-GAN further decreases the FID value for sketch synthesis and performs on par with the best single-stage instantiations. This gain indicates that the effect of these two techniques is at least partly additive.

Stacked refinement: Stacked cGANs dramatically decrease the FID of cGAN from 43.2 to 36.6 for sketch synthesis, and from 117.6 to 104.3 for photo synthesis. This suggests that stacked refinement is effective for improving the realism of synthesized images. Likewise, compared to CA-GAN, SCA-GAN further decreases FID by 3.5 (39.7 → 36.2) for sketch synthesis and by 10.7 (81.1 → 70.4) for photo synthesis. Besides, SCA-GAN achieves better results than stacked cGANs.

C. Face Sketch Synthesis

In this section, we compare the proposed method with a number of existing advanced methods, including MRF [16],
### Table I
Ablation study on the CUFS dataset. Our baseline is cGAN with a face photo/sketch as input. The best index in each column is shown in **boldface**. ↓ indicates lower is better, while ↑ higher is better.

<table>
<thead>
<tr>
<th>Remarks</th>
<th>Input Image</th>
<th>Input Composition</th>
<th>Compositional Loss</th>
<th>Stacked</th>
<th>Sketch Synthesis</th>
<th>Photo Synthesis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>FID↓ FSIM↑ NLDA↑</td>
<td>FID↓ FSIM↑ NLDA↑</td>
</tr>
<tr>
<td>cGAN (Base)</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>43.2 71.1 95.5</td>
<td>117.6 74.8 89.0</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>43.8 69.2 85.5</td>
<td>103.3 77.0 94.1</td>
</tr>
<tr>
<td>CA-GAN</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>40.5 71.3 95.2</td>
<td>81.1 78.0 98.8</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>-</td>
<td>✓</td>
<td>-</td>
<td>40.2 71.0 95.5</td>
<td>83.1 75.9 89.4</td>
</tr>
<tr>
<td>SCA-GAN</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>39.7 71.2 95.6</td>
<td>81.1 78.6 98.6</td>
</tr>
</tbody>
</table>

### Table II
Performance on face sketch-synthesis on the CUFS dataset and CUFSF dataset. The best and second best indices in each line are shown in **boldface** and **underline** format, respectively. ↓ indicates lower is better, while ↑ higher is better.

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Dataset</th>
<th>Traditional methods</th>
<th>Deep methods</th>
<th>(Deep) GANs based methods</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MRF</td>
<td>MWF</td>
<td>SSD</td>
</tr>
<tr>
<td>FID↓</td>
<td>CUFS</td>
<td>68.2</td>
<td>87.0</td>
<td>97.2</td>
</tr>
<tr>
<td></td>
<td>CUFSF</td>
<td>70.7</td>
<td>86.9</td>
<td>75.9</td>
</tr>
<tr>
<td></td>
<td>avg.</td>
<td>69.5</td>
<td>87.0</td>
<td>86.6</td>
</tr>
<tr>
<td>FSIM↑</td>
<td>CUFS</td>
<td>70.4</td>
<td>71.4</td>
<td>69.6</td>
</tr>
<tr>
<td></td>
<td>CUFSF</td>
<td>69.6</td>
<td>70.3</td>
<td>68.2</td>
</tr>
<tr>
<td></td>
<td>avg.</td>
<td>70.0</td>
<td>70.9</td>
<td>68.9</td>
</tr>
<tr>
<td>NLDA↑</td>
<td>CUFS</td>
<td>88.4</td>
<td>92.3</td>
<td>91.1</td>
</tr>
<tr>
<td></td>
<td>CUFSF</td>
<td>45.6</td>
<td>73.8</td>
<td>70.6</td>
</tr>
<tr>
<td></td>
<td>avg.</td>
<td>67.0</td>
<td>83.2</td>
<td>80.9</td>
</tr>
</tbody>
</table>

MWF [48], SSD [14], MrFSPS [17], RSLCR [1], MRNF [2], BFCN [49], DGFL [50], BP-GAN [4], and cGAN [3]. Synthesized images of existing methods are released by corresponding authors.

Table II show the results, where “avg.” denotes the average value of each criterion across the CUFS dataset and CUFSF dataset. Obviously, our final model, SCA-GAN, significantly decreases the previous state-of-the-art FID by a large margin across both datasets. Besides, CA-GAN obtains the second best FID values. This demonstrates that our methods dramatically improve the realism of the synthesized sketches, compared to existing methods. In addition, our methods are highly comparable with previous methods, in terms of FSIM and NLDA.

Fig. 3 presents some synthesized face sketches from different methods on the CUFS dataset and the CUFSF dataset. Due to space limitation, we only compare to several advanced methods here. Obviously, MrFSPS, RSLCR, and BFCN yield serious blurred effects and great deformation in various facial parts. In contrast, GANs based methods can generate sketch-like textures (e.g. hair region) and shadows. However, BP-GAN yields over-smooth sketch portraits, and cGAN yields deformations in synthesized sketches, especially in the mouth region. Notably, CA-GAN alleviates such artifacts, and SCA-GAN almost eliminates them.

To conclude, both the qualitative and quantitative evaluations shown in Table II and Fig. 3 demonstrate that both CA-GAN and SCA-GAN are capable of generating quality sketches. Specially, our methods achieve significantly gain in realism of synthesized sketches over previous state-of-the-art methods. Besides, our methods perform on par with previous state-of-art methods in term of FSIM and face sketch recognition accuracy.

### D. Face Photo Synthesis

We exchange the roles of the sketch and photo in the proposed model, and evaluate the face photo synthesis performance. To our best knowledge, only few methods have been proposed for face photo synthesis. Here we compare the proposed method with one advanced method: MrFSPS [17].

As shown in Table III, both CA-GAN and SCA-GAN achieves much lower FID values than MrFSPS and cGAN, while show slightly inferiority over MrFSPS in term of FSIM. Besides, our methods perform on par with cGAN in terms of both FID and NLDA. Nevertheless, our methods achieve the best balance between texture-realism and identity-consistency, and can well handle both the face sketch synthesis and photo synthesis tasks in a unified framework.

Fig. 4 shows examples of synthesized photos. Obviously, results of MrFSPS are heavily blurred. Besides, there is serious degradation in the synthesized photos by using cGAN. In contrast, the photos generated by CA-GAN or SCA-GAN consistently show considerable improvement in perceptual quality. Results of CA-GAN and SCA-GAN express more natural colors and details. Recall the quantitative evaluations shown in Table III, we can safely draw the conclusion that our methods are capable of generating natural face photos while preserving the identity of the input sketch.

1http://www.ihiworld.com/RSLCR.html
Fig. 3. Examples of synthesized face sketches on the the CUFS dataset and the CUFSF dataset. (a) Photo, (b) MrFSPS [3], (c) RSLCR[2], (d) FCN [1], (e) BP-GAN [8], (f) cGAN [4], (g) CA-GAN, (h) SCA-GAN, and (i) Sketch drawn by artist. From top to bottom, the examples are selected from the CUHK student dataset [38], the AR dataset [39], the XM2VTS dataset [40], and the CUFSF dataset [41], sequentially.

TABLE III

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Dataset</th>
<th>MrFSPS</th>
<th>cGAN</th>
<th>CA-GAN</th>
<th>SCA-GAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>FID↓</td>
<td>CUFS</td>
<td>92.0</td>
<td>88.7</td>
<td>83.9</td>
<td>70.4</td>
</tr>
<tr>
<td></td>
<td>CUFSF</td>
<td>95.6</td>
<td>33.1</td>
<td>32.1</td>
<td>32.4</td>
</tr>
<tr>
<td></td>
<td>avg.</td>
<td>93.8</td>
<td>60.9</td>
<td>58.0</td>
<td>51.4</td>
</tr>
<tr>
<td>FSIM↑</td>
<td>CUFS</td>
<td>80.3</td>
<td>76.2</td>
<td>76.5</td>
<td>78.6</td>
</tr>
<tr>
<td></td>
<td>CUFSF</td>
<td>79.3</td>
<td>79.5</td>
<td>79.1</td>
<td>79.7</td>
</tr>
<tr>
<td></td>
<td>avg.</td>
<td>79.8</td>
<td>77.8</td>
<td>77.8</td>
<td>78.7</td>
</tr>
<tr>
<td>NLDA↑</td>
<td>CUFS</td>
<td>96.7</td>
<td>94.8</td>
<td>98.6</td>
<td>99.2</td>
</tr>
<tr>
<td></td>
<td>CUFSF</td>
<td>59.4</td>
<td>77.5</td>
<td>73.4</td>
<td>74.9</td>
</tr>
<tr>
<td></td>
<td>avg.</td>
<td>78.2</td>
<td>86.1</td>
<td>86.0</td>
<td>87.1</td>
</tr>
</tbody>
</table>

E. Robustness Evaluation

To verify the generalization ability of the learned model, we apply the model trained on the CUFS training dataset to faces in the wild.

1) Lighting and Pose Variations: We first apply the learned models to a set of face photos with lighting variation and pose variation. Fig. 5 shows some synthesized results from cGAN, CA-GAN, and SCA-GAN. Clearly, results of cGAN exists blurring and inky artifacts over dark regions. In contrast, the photos synthesized by using CA-GAN and SCA-GAN show less artifacts and express improved details over eye and mouse regions. Additionally, results of SCA-GAN show the best quality.

2) Face photo-sketch synthesis of national celebrities: We further test the learned models on the photos and sketches of national celebrities. These photos and sketches are downloaded from the web, and contain different lighting conditions and backgrounds compared with the images in the training set. Fig. 6 and Fig. 7 shows the synthesized sketches and photos, respectively. Obviously, our results express more natural textures and details than cGAN, for both sketch synthesis and photo synthesis. Both CA-GAN and SCA-GAN show outstanding generalization ability in the sketch synthesis task.

The synthesized photos here are dissatisfactory. This might
Fig. 4. Examples of synthesized face photos. (a) Sketch drawn by artist, (b) MrFSPS [17], (c) cGAN, (d) CA-GAN, (e) SCA-GAN, and (f) ground-truth photo. From top to bottom, the examples are selected from the CUHK student dataset [38], the AR dataset [39], the XM2VTS dataset [40], and the CUFSF dataset [41], sequentially.

Fig. 5. Robustness to lighting and pose variations. (a) Photo, (b) cGAN, (c) CA-GAN, and (d) SCA-GAN.

Fig. 6. Synthesized sketches of national celebrities. (a) Photo, (b) cGAN, (c) CA-GAN, and (d) SCA-GAN.

Fig. 7. Synthesized photos of national celebrities. (a) Sketch, (b) cGAN, (c) CA-GAN, and (d) SCA-GAN.
be due to the great divergence between the input sketches in terms of textures and styles. It is necessary to further improve the generalization ability of the photo synthesis models.

F. Stability of the Training Procedure

We empirically find that our proposed approaches considerably stabilizes the training procedure of the network. Fig. 8 shows the (smoothed) training loss curves related to cGAN [3], CA-GAN, and SCA-GAN on the CUFS dataset. Specially, Fig. 8(a) and Fig. 8(b) shows the reconstruction error (Global $L_1$ loss) and the adversarial loss in the sketch synthesis task; Fig. 8(c) and Fig. 8(d) show the reconstruction error and the adversarial loss in the photo synthesis task, respectively. For clarity, we smooth the initial loss curves by averaging adjacent 40 loss values.

Obviously, there are large impulses in the adversarial loss of cGAN. In contrast, losses related to both CA-GAN and SCA-GAN are much smoother. The reconstruction error of both CA-GAN and SCA-GAN are aslo smaller than that of cGAN. Additionally, SCA-GAN achieves the least reconstruction errors and smoothest loss curves. This observation demonstrates that the proposed compositional loss both speeds the training up and greatly stabilizes it.

V. Conclusion

In this paper, we propose a novel composition-aided generative adversarial network (CA-GAN) for face photo-sketch synthesis. Our approach dramatically improves the realism of the synthesized face photos and sketches over previous state-of-the-art methods. We hope that the presented approach can support applications of other image generation problems. Besides, it is essential to develop models that can handle photos/sketches with great variations in head poses, lighting conditions, and styles. Finally, exciting work remains to be done to qualitatively evaluate the quality of synthesized sketches and photos.

REFERENCES

Fig. 8. Training loss curves of cGAN, CA-GAN, and SCA-GAN, on the CUFS dataset. (a) Reconstruction error in the sketch synthesis task, (b) adversarial loss in the sketch synthesis task, (c) reconstruction error in the photo synthesis task, and (d) adversarial loss in the photo synthesis task.


