# Improving Face Sketch Recognition via Adversarial Sketch-Photo Transformation

Shikang Yu[1,3], Hu Han[1,2,*], Shiguang Shan[1,2,3,4], Antitza Dantcheva[5], and Xilin Chen[1,3]
[1] Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
[2] Peng Cheng Laboratory, Shenzhen, China
[3] University of Chinese Academy of Sciences, Beijing 100049, China
[4] CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai, China
[5] Inria, Sophia Antipolis, France
shikang.yu@vipl.ict.ac.cn; {hanhu, sgshan, xlchen}@ict.ac.cn; antitza.dantcheva@inria.fr

***Abstract*— Face sketch-photo transformation has broad applications in forensics, law enforcement, and digital entertainment, particular for face recognition systems that are designed for photo-to-photo matching. While there are a number of methods for face photo-to-sketch transformation, studies on sketch-to-photo transformation remain limited. In this paper, we propose a novel conditional CycleGAN for face sketch-to-photo transformation. Specifically, we leverage the advantages of CycleGAN and conditional GANs and design a feature-level loss to assure the high quality of the generated face photos from sketches. The generated face photos are used, as a replacement of face sketches, and particularly for face identification against a gallery set of mugshot photos. Experimental results on the public-domain database CUFSF show that the proposed approach is able to generate realistic photos from sketches, and the generated photos are instrumental in improving the sketch identification accuracy against a large gallery set.**

## I. Introduction

Facial sketches are widely used by law enforcement agencies as support for suspect identification. Typically, forensic sketches are hand drawn based on verbal descriptions of the eyewitness or the victim. In addition to circulation of such forensic sketches on TV and newspapers, the police departments tend to utilize face sketches as probes, in automated face recognition searches in mugshot databases. While face sketch-photo transformation can be instrumental for the improvement of such sketch based face recognition, it is also of interest for entertainment. For example, in social media, many people use face sketches as their avatar figures on Facebook, Twitter, QQ, and Wechat.

However, given the heterogeneous nature of photos and sketches stemming from different generation mechanisms (i.e., intensity by digital sensor vs. line drawing by hand), there can be large geometric deformations and texture differences between a face photo and its associated sketch. These factors make sketch based face recognition a challenging heterogeneous face recognition (HFR) problem.

Towards tackling this problem many methods have been proposed that can be classified into two categories: (i) photo-sketch transformation [1]–[6], and (ii) modality-invariant feature learning [7]–[10]. The benefit of the former category relates to the conversion of sketches into the same modality as photos, and hence lies in the ability to utilize existing photo-based face recognition methods. Thus, the applicability of the existing photo-based face recognition algorithms can be greatly expanded.

Current methods for face photo-sketch transformation can be mainly grouped into example-based methods and regression-based methods. *Example-based methods* assume that the corresponding sketches (or patches of sketches) of two similar face photos (or patches of face photos) are also similar. Such methods rely on face photo-sketch pairs in the training set to synthesize images. In order to achieve good transformation results, these methods usually require a large number of photo-sketch pairs. However, the computational cost may also grow linearly with the increase of the training set size. *Regression-based methods* overcome the issues mentioned above and the most time-consuming part only exists in the training stage when learning the mapping between face photos and sketches, but the inference/testing stage can be fast. In this paper, we propose a Generative Adversarial Network (GAN) for face sketch-to-photo transformation, leveraging the advantages of CycleGAN [11] and conditional GANs [12]. We have designed a new feature-level loss, which is jointly used with the traditional image-level adversarial loss to ensure the quality of the synthesized photos. The proposed approach outperforms state-of-the-art approaches for synthesizing photos in terms of structural similarity index (SSIM). More importantly, the synthesized photos of our approach are found to be more instrumental in improving the sketch-to-photo matching accuracy.

The rest of this paper is organized as follows: Section II summarizes representative methods of face photo-to-sketch transformation, and GANs. Section III provides details of the proposed method and the designed feature-level loss. Experimental results and analysis are presented in Section IV. Finally, we conclude this work in Section V.

* Corresponding author.

978-1-7281-0089-0/19/$31.00 

## II. Related Work

### A. Face Photo-Sketch Synthesis

Notable approaches in face sketch synthesis include the works [1]–[5], [13]–[19]. Tang and Wang [1], [2] performed photo-sketch transformation under the assumption that the transformation can be viewed as a linear mapping. They used principal component analysis (PCA) [20] to reconstruct the input photo from training photos and then synthesize the sketch by a linear combination of corresponding sketches weighted by the same reconstruction coefficients. We note that the whole face and whole sketch cannot be well represented by only a simple linear transformation, specifically the high-frequency information of the generated sketches is not well synthesized, and hence the output sketch tends to be blurry. Considering the restrained presentation ability of linear transformation, Liu et al. [3] and Liu et al. [4] presented nonlinear transformation methods for face sketch synthesis by exploiting local linear embedding (LLE) [21]. This constituted a classical manifold learning method, which adopted the patch-based strategy. A test photo and the photo-sketch pairs in the training set were divided into several overlapping image patches. K-NN was performed to search the $K$ nearest photo patches in the training set, and the obtained corresponding $K$ sketch patches were used to synthesize the output sketch patch. The output sketch is a weighted combination of these $K$ sketch patches using the same weights. This method outperformed former approaches, has though other limitations, as it does not consider the compatibility or relationship between neighboring patches, i.e., each patch was synthesized independently. Considering such a relation, Wang and Tang [5] employed the Markov Random Fields (MRF) to model the distribution patches, i.e., the distribution between a test photo patch and the nearest photo patches, as well as the distribution between adjacent synthesized sketch patches. However, such a method cannot generate a new image patch that does not exist in the training photo-sketch pairs, and it is an NP-hard to optimize. Considering that sketches generated by Liu et al. [3] were noisy, Song et al. [13] presented a real-time example-based face sketch synthesis method which considers face sketch synthesis (FSS) as an image denoising problem and the method can be speeded up, when processed with a GPU. The synthesized sketches looked flat and blunted due to denoising by average operators.

Sparse coding [22] and dictionary learning are prevalent in image reconstructing [23]–[26], and they have been utilized in sketch synthesis from photos [14]–[16]. Chang et al. [14] applied sparse representation to face sketch synthesis, where the training photo-sketch set was firstly divided into overlapping regions and then a local coupled dictionary was constructed using sparse coding. For each region in a test photo, its sparse representation coefficients were computed to determine the corresponding sketch patches for reconstruction. Additionally, a local smoothness constraint was enforced between overlapping patches.

*Regression-based methods* aim to learn a regression model to transform a face photo to a face sketch or vice versa. Zhu and Wang [17] used a Ridge Regression model to learn the mapping between photo patches and their corresponding sketch patches. Zhang et al. [18] used a convolution network (CNN) as regression model. However, the generated sketches still looked blurry due to the limited representation ability of their loss function (mean square error, MSE). Zhu et al. [19] combined a generative example-based method with discriminatively trained deep convolutional neural networks for face sketch synthesis. Herein the procedures in the test stage were complicated, which not only included a single transformation by a network (thus the associated computation cost was not low) and it was only able to transform face photos into sketches in one direction.

### B. Generative Adversarial Networks (GANs)

GANs were introduced by Goodfellow et al. [27] and comprise a generative model $G$ and a discriminative model $D$, jointly trained in an adversarial way. $G$ generates data with distribution resembling the training data, while D estimates the probability that a sample comes from the training data rather than from $G$. Both models are multilayer perceptrons in general, and their abilities are poor in the beginning of the training. The two models compete against each other, resulting in their abilities improving so the generative model $G$ can produce data with a similar distribution to the training data. GANs obtain impressive results without any complicated process. Since the advent of GANs, numerous variations have been proposed. Conditional GANs (cGANs) [12] added an auxiliary information $y$ into the original concept, which relates to a modality or class labels. By adding extra information, the generation process can be directed. Least Squares GANs (LSGANs) [28] replaced the sigmoid cross entropy loss function from the original GANs with least squares loss function for the discriminator, which tackled the vanishing gradient problem during the learning process. Deviating from cGANs, Pix2Pix [29] used U-Net [30] as the generator network and proposed "PatchGAN" classifier as discriminator accountable for the image-to-image transformation. L1 distance term was added to the objective function to capture the low-frequency information. The U-Net structure preserves a part of important information of the input image and directs it to the output image using skip connections. CycleGAN [11] adopted the cycle-consistency loss to break the limitations of training image-to-image transformation, while paired training data is not available. It trained four neural networks: a generator network for translating images from domain $A$ to domain $B$, an additional one for translating images from domain $B$ to domain $A$, a discriminator network for estimating the probability that the image is from domain $B$ rather than $A$, and finally one performing the opposite estimation. Unlike neural style transfer [31], it performed transformation, mimicking the style on the entire image collection rather than a single selected image.

The here proposed approach is motived by cGANs [12], Pix2Pix [29] and CycleGAN [11], but it differs from these

methods in that: (i) we propose a novel feature-level loss to assure the synthesized photo quality and its effectiveness for face recognition, (ii) we employ auxiliary information during network training to improve the robustness of the model. We note that the network inference does not require such auxiliary information.

## III. Proposed Approach

As shown in Fig. 1, our approach learns two mappings between sketch domain $S$ and photo domain $P$, i.e., $S \rightarrow P$, a mapping from sketch domain to photo domain, as well as the opposite mapping $P \rightarrow S$. In order to achieve this, we use two generator networks $G_{S2P}$ and $G_{P2S}$, of which $G_{S2P}$ aims to generate the corresponding photos from input face sketches, and $G_{P2S}$ operates symmetrically. We also train two discriminator networks $D_P$ and $D_S$. $D_P$ takes the input of two images, i.e., a synthesized face photo $U$ generated by $G_{S2P}(X)$ and the corresponding real face sketch $X$, and the output $D_P(X, U)$ denotes the probability that the synthetic face photo $U$ is likely to be the real photo corresponding to the real sketch $X$. Similarly, $D_S(Y, V)$ gives the probability that a synthesized face sketch $V$ is likely to be the real sketch corresponding to the input photo $Y$.

Motivated by conditional GANs, in which both the generator and discriminator are provided with auxiliary information to enhance the network training, the discriminators of our network also take a synthesized face image and a real image from the other modality as input, in which the real image is used as the auxiliary information. In addition, we design a feature-level loss to penalize the differences between a synthesized photo and the ground-truth photo in the feature space. Thus, it works as a complementary loss to the L1 loss to enhance the ability of generator $G_{S2P}$. The ablation study in our experiments below shows the effectiveness of our feature-level loss.

### A. Network Architecture

We use the same network architecture for the two generators $G_{S2P}$ and $G_{P2S}$. Similarly, the two discriminators $D_S$ and $D_P$ share the same network architecture. In the following, we proceed to provide details on the generator and discriminator architectures.

We use following notations to facilitate our description below: C(k,n) denotes a Convolution-BatchNorm-Relu layer with $k$ filters and $n$-channels in the output. CD(k,n) denotes a Convolution-Batchnorm-Relu-Dropout layer with $k$ filters and $n$-channels in the output, the dropout rate is 50%. All convolutions are $4 \times 4$ spatial filters with a stride of 2.

*1) Generator:* We use a U-Net [30] structure to build our generator because of its success in Pix2Pix, which was utilized in the context of image transformation [29]. One advantage of the U-Net is that it is able to preserve a part of the important information from the input image using skip connections. Such a property can be useful for our sketch-to-photo transformation task, because the synthesized face photo will be used for face identification, and the retained information from the input sketch image may retail more subject discriminative information.

Specifically, the encoder of our generator has six convolutional layers: C(64,64)-C(128,128)-C(256,256)-C(512,512)-C(512,512)-C(512,512), and the decoder of our generator has six convolutional layers as well: CD(512,1024)-CD(512,1024)-CD(256,512)-CD(128,256)-CD(64,128)-CD(3,3). Since the face sketch images are single-channel gray-scale images, we duplicate the images into three channels, and use them as the input of the generator network. Such a three-channel duplication operation also allows for a compatibility with sketch-to-color-photo transformation, which we intend to study in a future work.

We do use batch normalization in both the encoder and decoder, but it is not applied in the first and last layers of the encoder (i.e., C(64,64) and C(512,512)), and the last layer in the decoder (i.e., CD(3,3)). All ReLus in the encoder except the last layer are leaky with slope 0.2, and all the other ReLus in the generator are not leaky. There is a Tanh in the last layer CD(3,3) in the decoder instead of Relu. The skip connections concatenate activations from layer $i$ in the encoder to layer $6 - i$ in the decoder, which leads to the number of the output channels of layers 1-5 in the decoder to be twice the number of its filters.

In order to have a better performance of sketch-to-photo transformation, we tested a more powerful sketch-to-photo generator $G_{S2P}$, with more layers (two or four convolution layers). However, our experiments show that it does not improve the performance. This might be due to the small size of the training set.

*2) Discriminator:* Our discriminator consists of three convolution layers (see Table I). Again the input to our discriminator is a duplicated three-channel image.

TABLE I
The architecture of the discriminator in our approach.

| layer no | output size | Description |
|---|---|---|
| i | $64 \times 64$<br>3 channels | the gray-scale images are<br>duplicated as 3 channels |
| d1 | $64 \times 64$<br>64 channels | conv $1 \times 1$, 64 filters, stride=1,<br>(padding=0, use_bias=True)<br>LeakyReLu(0.2) |
| d2 | $64 \times 64$<br>128 channels | conv $1 \times 1$, 128 filters, stride=1,<br>(padding=0, use_bias=False)<br>BatchNorm<br>LeakyReLu(0.2) |
| d3 | $64 \times 64$<br>1 channels | conv $1 \times 1$, 1 filters, stride=1,<br>(padding=0, use_bias=False) |

### B. Objective Function

The whole network comprises of five pertinent losses: (1) adversarial loss, (2) L1 loss, (3) cycle-consistency loss, (4) identity mapping loss, and (5) face feature loss, which we proceed to describe.

*1) **Adversarial Loss**:* The design of the adversarial loss is an important part of the objective, as it decides whether the synthesized photo (or sketch) can be as close to the real photo (or real sketch) as possible. Two adversarial losses

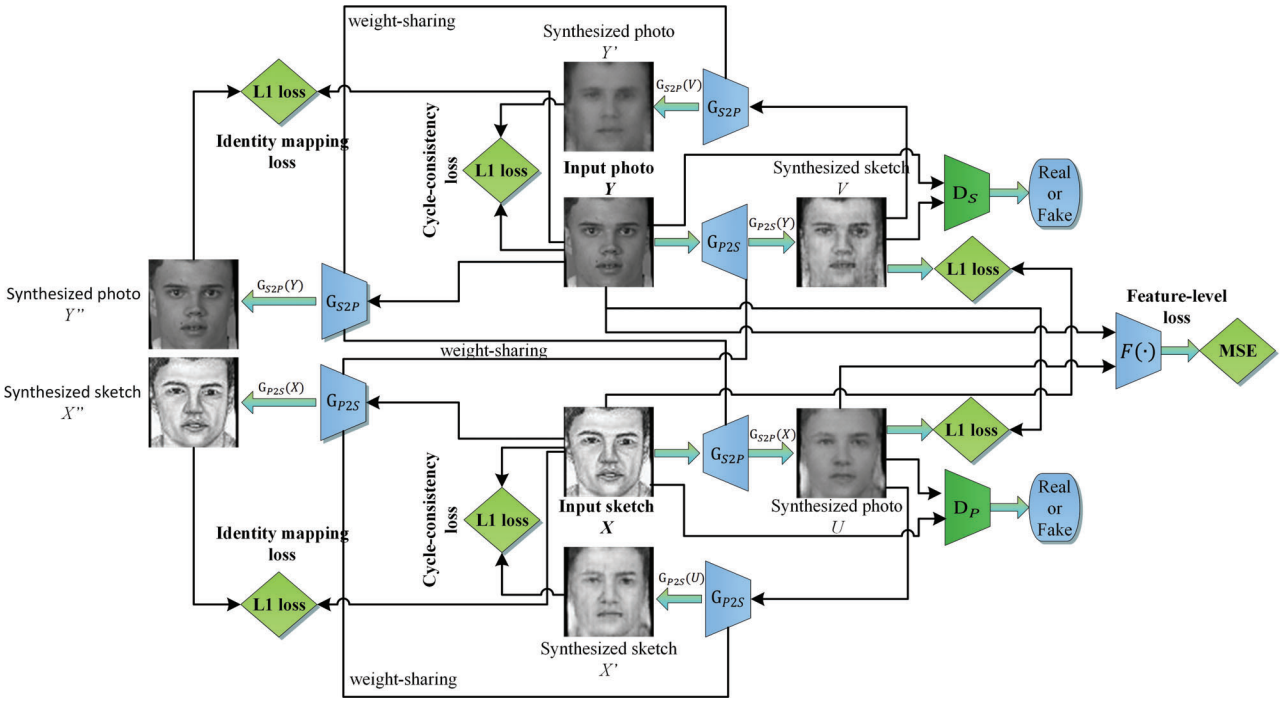


Fig. 1. Overview of the proposed GAN for sketch-to-photo transformation using feature-level loss.

used in $G_{S2P}, D_P, G_{P2S}$, and $D_S$ are defined as follows

$$\begin{aligned}\mathcal{L}_{\text{S2P}}(G_{S2P}, D_P) = &\text{E}_{x,y\sim p_{data}(x,y)}[\log D_P(x,y)] \\ &+ \text{E}_{x\sim p_{data}(x)}[\log(1 - D_P(x, G_{S2P}(x)))],\end{aligned} \tag{1}$$

$$\begin{aligned}\mathcal{L}_{\text{P2S}}(G_{P2S}, D_S) = &\text{E}_{x,y\sim p_{data}(x,y)}[\log D_S(y,x)] \\ &+ \text{E}_{y\sim p_{data}(y)}[\log(1 - D_S(y, G_{P2S}(y)))].\end{aligned} \tag{2}$$

Thus, $\mathcal{L}_{S2P}$ is accountable to train the generator $G_{S2P}$ and the discriminator $D_P$ in an adversarial way. In particular, $G_{S2P}$ is expected to generate face photos as similar to the real sketches as possible, so that $D_P$ is not able to differentiate the synthesized photos from real photos. At the same time, the discriminator $D_P$ aims to distinguish between the real photos and the synthesized photos, i.e., $\min_{G_{S2P}} \max_{D_P} \mathcal{L}_{\text{S2P}}(G_{S2P}, D_P)$.

Instead of using the original form of adversarial loss, we use the least squares loss in LSGANs [28] as the $\mathcal{L}_{\text{S2P}}(G_{S2P}, D_P)$, i.e,

$$\begin{aligned}\min_{D_P} = &\frac{1}{2}\text{E}_{x\sim p_{data}(x)}[(D_P(x, G_{S2P}(x)))^2] \\ &+ \frac{1}{2}\text{E}_{x,y\sim p_{data}(x,y)}[(D_P(x,y) - 1)^2] \\ \min_{G_{S2P}} = &\text{E}_{x\sim p_{data}(x)}[(D_P(x, G_{S2P}(x)) - 1)^2].\end{aligned} \tag{3}$$

Similarly, we define $\mathcal{L}_{\text{P2S}}(G_{P2S}, D_S)$ as

$$\begin{aligned}\min_{D_S} = &\frac{1}{2}\text{E}_{y\sim p_{data}(y)}[(D_S(y, G_{P2S}(y)))^2] \\ &+ \frac{1}{2}\text{E}_{x,y\sim p_{data}(x,y)}[(D_S(y,x) - 1)^2] \\ \min_{G_{P2S}} = &\text{E}_{y\sim p_{data}(y)}[(D_S(y, G_{P2S}(y)) - 1)^2].\end{aligned} \tag{4}$$

We find that using these adversarial losses leads to better results in terms of the face identification rate.

*2) **L1 Loss**:* We use a L1 loss to penalize the difference between the synthesized photos (or sketches) and the ground-truth photos (or ground-truth sketches).

$$\mathcal{L}_{\text{L1}}(G_{S2P}) = \text{E}_{x,y\sim p_{data}(x,y)}[||y - G_{S2P}(x)||_1] \tag{5}$$

and

$$\mathcal{L}_{\text{L1}}(G_{P2S}) = \text{E}_{x,y\sim p_{data}(x,y)}[||x - G_{P2S}(y)||_1]. \tag{6}$$

With the L1 loss the two generators $G_{S2P}$ and $G_{P2S}$ are expected to generate synthesized photos and synthesized sketches that are as similar to the ground-truth photos and the ground-truth sketches as possible.

*3) **Cycle-Consistency Loss**:* As shown in [11], the cycle-consistency loss requires the image translation cycle to be able to bring an input $x$ back to its original domain, i.e., $G_{P2S}(G_{S2P}(x)) \approx x$ (see Fig. 1). Cycle-consistency loss is particularly useful when no paired training data is available. Even though our training data (sketch and photo) is paired, we also use the cycle-consistency loss to enhance the sketch-photo translation capability of the whole network.

$$\begin{aligned}\mathcal{L}_{\text{cyc}}(G_{S2P}, G_{P2S}) = &\text{E}_{x\sim p_{data}(x)}[||G_{P2S}(G_{S2P}(x)) - x||_1] \\ &+ \text{E}_{y\sim p_{data}(y)}[||G_{S2P}(G_{P2S}(y)) - y||_1]\end{aligned} \tag{7}$$

*4) **Identity Mapping Loss**:* The three losses above (adversarial loss, L1 loss and cycle-consistency loss) emphasize that the two generators ($G_{S2P}$ and $G_{P2S}$) can generate high-quality synthesized photos and sketches w.r.t. the ground-truth photos and sketches. We also expect that if a photo

image is input to the sketch-to-photo generator $G_{S2P}$, the output should remain the input photo. It is similar for the photo-to-sketch generator $G_{P2S}$. Such a constraint is meant to enforce the two generators to learn the essential characteristics of the photo images and the sketch images. Similar to Zhu et al. [11], we use a identity-mapping loss to represent such a constraint

$$\begin{aligned}\mathcal{L}_{\text{idt}}(G_{S2P}, G_{P2S}) = \mathrm{E}_{y\sim p_{data}(y)}[||G_{S2P}(y) - y||_1] \\ + \mathrm{E}_{x\sim p_{data}(x)}[||G_{P2S}(x) - x||_1].\end{aligned} \tag{8}$$

*5) **Face Feature Loss**:* All four above described losses operate at the image level, in order to assist in the network training. Motivated by the success of the perceptual loss for image style transfer and super-resolution [32], we design a feature-level loss on the feature domain to further enhance the network learning. Specifically, we use a pre-trained face classification network (i.e., Light CNN-29 v2 model [33]) to extract features from the synthesized photo and the ground-truth photo images, which are then used for compute the MSE loss

$$\begin{aligned}&\mathcal{L}_{\text{feature}}(G_{S2P}) = \\ &\mathrm{E}_{x,y\sim p_{data}(x,y)} \sum_{i=1}^{4} \text{MSE}[F_i(y) - F_i(G_{S2P}(x))],\end{aligned} \tag{9}$$

where $F_i(x)$ represents the deep features of the convolutional layer before the $i$-th pooling layer extracted for $x$ by the pre-trained network $F$. The real photo $y$ and the synthesized photo $G_{S2P}(x)$ are resized to $128 \times 128$ before extracting features using the pre-trained Light CNN-29 v2 [33], which accepts input images of $128 \times 128$. Besides improving the synthesized photo quality, the proposed feature-level loss is additionally expected to assure that the synthesized photos from sketches can be used to improve the face identification accuracy of the widely used deep face recognition systems for photo-to-photo matching (see Fig. 2).

With the definition of the above five losses, the full objective function can be written as

$$\begin{aligned}&\mathcal{L}(G_{S2P}, G_{P2S}, D_P, D_S) \\ &\quad= \mathcal{L}_{\text{S2P}}(G_{S2P}, D_P) + \mathcal{L}_{\text{P2S}}(G_{P2S}, D_S) \\ &\quad+ \mathcal{L}_{\text{feature}}(G_{S2P}) + \lambda\mathcal{L}_{\text{cyc}}(G_{S2P}, G_{P2S}) \\ &\quad+ \lambda_{\text{L1}}\mathcal{L}_{\text{L1}}(G_{S2P}) + \lambda_{\text{L1}}\mathcal{L}_{\text{L1}}(G_{P2S}) \\ &\quad+ \lambda \times \lambda_{\text{idt}}\mathcal{L}_{\text{idt}}(G_{S2P}, G_{P2S}).\end{aligned} \tag{10}$$

We empirically set $\lambda = 10$, $\lambda_{\text{idt}} = 0.5$ (suggested in CycleGAN [11]), and $\lambda_{\text{L1}} = 100$ in our experiments. We notice that the final face identification performance using the synthesized photos from sketches is sensitive to $\lambda_{\text{L1}}$. Using $\lambda_{\text{L1}} = 100$ achieves relatively higher face identification performance. The identification rate will degrade if we use a too small $\lambda_{\text{L1}}$ or a too large $\lambda_{\text{L1}}$. We notice the possible reason is that if we set $\lambda_{\text{L1}}$ smaller than 100, the generated photos tend to be noisy, and if we set $\lambda_{\text{L1}}$ larger than 100, the generated photos tend to be over-smoothed.

### C. Network Training

We use a pre-trained Light CNN-29 v2 model [33] with 1,094 neurons in the last fully connection layer as the feature extraction network $F$ in our face feature loss. This network is fine-tuned offline using the sketch-photo data in the training set of CUFSF. We use an initial learning rate $\alpha_0 = 0.00001$, and train 80 epochs in total. The learning rate decays in the form of $\alpha = \alpha_0 \times scale^{\lfloor \frac{epoch-1}{10} \rfloor}$, in which $scale = 0.457305$.

We then train our network using an initial learning rate $\alpha_0 = 0.0002$, and $batchsize = 1$. We train the individual modules of the entire network for 200 epochs in total. The entire network uses the same learning rate, which decays in the form of $\alpha = \alpha_0 \times [1 - \max(0, \frac{epoch+1-100}{101})]$. Finally, we choose the model with the best identification rate for evaluations.

## IV. Experimental Evaluations

### A. Dataset and Baselines

The CUHK Face Sketch FERET Database (CUFSF) [6] is a widely used database for studying face photo-sketch synthesis and face sketch recognition. It includes 1,194 face photos and 1,194 corresponding face sketches drawn by artists. For each photo or sketch, we first detect the five facial keypoints (two midpoints of eyes, a nose tip, two mouth corners) using an open source face recognition SDK—SeetaFace[1], and then align and crop the faces based on the facial keypoints. If the facial keypoints cannot be detected, we use manually annotated keypoints. The cropped face images are resized to $64 \times 64$ for network training. We randomly choose 100 photo-sketch pairs as the testing set, and the remaining photo-sketch pairs are used for training. While there are some other face sketch databases, e.g., PRIP-Viewed Sketch [8], [34], IIT-D sketch [35], the sizes of these database are very small (about 200 sketch-photo pairs); thus we choose to report the results on the relative larger CUFSF database (more than 1,000 sketch-photo pairs).

After generating photos from the probe sketches, as shown in Fig. 2, we use them to replace the sketches to perform face identification. We expect the generated photos by our approach to be able to improve the face recognition performance. While there are a number of face recognition models available, the open source SeetaFace reported 97.1% accuracy on the LFW database [36] and so it constitutes a strong baseline face matcher for our following experiments. The probe set contains 100 synthesized photos generated by our approach and the gallery set consists of the 100 mated face photos, as well as 10,000 photos from the MORPH database [37] as background gallery face image to populate the gallery set. We compare the proposed method with a number of state-of-the-art approaches that can perform sketch-to-photo transformation such as CycleGAN [11] and Pix2Pix [29]. There are some variations of CycleGAN, such as conditional CycleGAN [38], which utilizes face attributes (e.g., skin color or hair color) as the auxiliary information to improve

[1] https://github.com/seetaface/SeetaFaceEngine

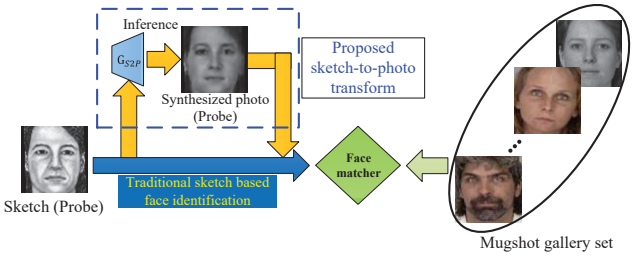


Fig. 2. Diagram of using the synthesized photo by our approach to perform face identification.

the network training. However, in our experiments, conditional CycleGAN needs extra information when transforming sketches to photos, and does not show advantage in sketch-to-photo transformation and the face identification tasks. Hence we choose to report the performance of CycleGAN.

### B. Evaluation Metrics

Besides the subjective comparisons of synthesized photos, we use two quantitative evaluation measures: Cumulative Match Characteristic (CMC) curve widely used in face identification experiment, and structural similarity index (SSIM) [39], widely used for image quality assessment[2].

### C. Results

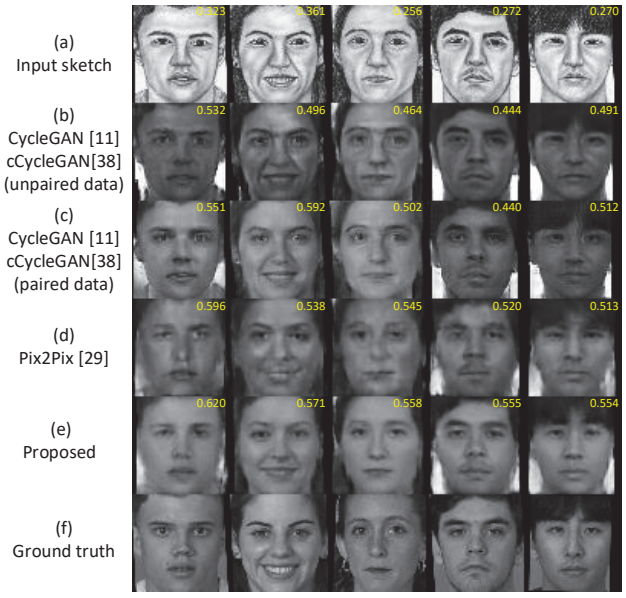


Fig. 3. Synthesized face photos by different approaches. (a) input face sketches from CUFSF, (b-c) synthesized face photos by CycleGAN [11] using paired and unpaired data respectively, (d) synthesized face photos by Pix2Pix [29], (e) synthesized face photos by the proposed method, (f) the ground-truth photos. The number at the upper-right corner of each image is the SSIM score of that image w.r.t. the ground-truth photo.

Fig. 3 shows example photos generated by the proposed approach, and a number of state-of-the-art approaches. We evaluate CycleGAN using both, paired and unpaired sketch-photo images, as it can work with either paired or unpaired data. We can see that CycleGAN fails to produce photo-realistic results, and the synthesized photos often contain dark shades. Compared with CycleGAN, both Pix2Pix and our method generate more visually pleasing results. However, some deformations can be observed in the face photos generated by Pix2Pix, while such deformations do not exist in the synthesized photos by our approach. The face identification accuracies and SSIM scores by our approach outperform those of the state-of-the-art methods. These results suggest that the proposed approach is able to convert sketches to photos with better image quality.

[2]We did not report the Inception Score [40] because it is designed for evaluating multi-class image generation task; however, we aim to improve face identification accuracy via face sketch-to-photo transformation.

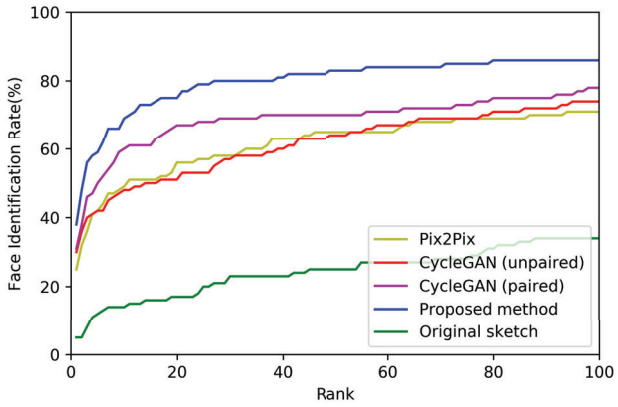


Fig. 4. CMC curves of face identification using the synthesized photos by the proposed method, Pix2Pix, and CycleGAN respectively. We also give the CMC curve of using the original face sketches for face identification.

*1) Face identification using synthesized photos:* We perform face identification experiments using the synthesized photos by different methods to see whether they are able to improve face identification accuracy. Fig. 4 shows the CMC curves of individual methods. As expected, directly matching the probe sketches to the gallery photos offers the lowest matching rate. The synthesized photos by Pix2Pix, CycleGAN, and the proposed method lead to a significantly improved accuracy as opposed to using the probe sketches (see Fig. 4). However, the proposed method outperforms the state-of-the-art methods from rank-1 to rank-100 by a large margin (7% at rank-1). For easy comparison, we also report the rank-1 to rank-5 face identification accuracies using the synthesized photos by all approaches in Table II. Sketch to photo matching is a challenging heterogeneous face recognition task, because of the big modality gap. Given the limited training data (about 1,000 pairs), and the large gallery set with 10K background images, we here note that 7% at rank-1 improvement against the state-of-the-art methods is encouraging. In addition, the designed joint loss function does not increase the computational cost in network inference, and is only used during network training. These results suggest that the proposed approach is effective in reducing the modality gap between sketches and photos, and instrumental in improving the face sketch recognition performance.

The main reason why the proposed sketch-to-photo trans-

TABLE II

THE RANK-1 TO RANK-5 FACE IDENTIFICATION RATES USING THE SYNTHESIZED PHOTOS GENERATED BY DIFFERENT APPROACHES.

| Approach | Rank 1 | Rank 2 | Rank 3 | Rank 4 | Rank 5 |
|---|---|---|---|---|---|
| Original sketch | 5% | 5% | 8% | 11% | 12% |
| CycleGAN (unpaired) | 30% | 36% | 40% | 41% | 42% |
| CycleGAN (paired) | 31% | 38% | 46% | 47% | 50% |
| Pix2Pix | 25% | 32% | 36% | 41% | 42% |
| Proposed | **38%** | **48%** | **56%** | **58%** | **59%** |

TABLE III

THE AVERAGE SSIM BY DIFFERENT APPROACHES FOR SKETCH-TO-PHOTO TRANSFORMATION ON CUFSF.

| Approach | SSIM |
|---|---|
| Original sketch | 0.2836 |
| CycleGAN (unpaired) | 0.4796 |
| CycleGAN (paired) | 0.5174 |
| Pix2Pix | 0.5476 |
| Proposed | **0.5517** |

formation approach is effective in improving the face identification accuracy is that while most existing approaches mainly focus on improving visual quality of synthesized images [32], [41], our approach focuses on obtaining high-quality synthesized photos that retain as much identity information as possible, i.e., through the L1 loss, cycle-consistency loss and the feature-level loss.

*2) Image quality of synthesized photos:* The average SSIM of different approaches are reported in Table III. We observe that the proposed method achieves the highest SSIM among all methods. This again suggests that the proposed approach performs well in synthesizing photos from input face sketches, i.e., useful for law enforcement in converting the face sketch of a suspect into a photo, which can be a pre-processing step for further distribution via TV, social media or newspaper.

### D. Ablation Study

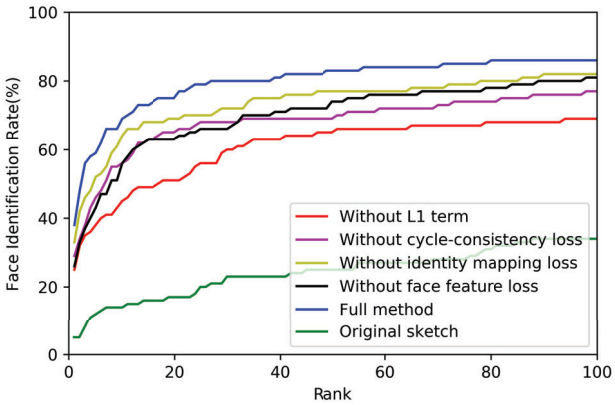


Fig. 5. Face identification performance (in terms of CMC) using the synthesized face images by the proposed approach under ablation study.

The objective function of our approach contains multiple losses, and we provide here an ablation study to verify the effectiveness of individual parts, such as the L1 term, cycle-consistency loss, identity mapping loss, and feature-level loss. We drop one component each time during the ablation study and train the ablated models under the same settings as the whole method. We generate photos from sketches using four ablated models, and again use the synthesized photos for face identification experiments. As shown in Fig. 5, removing identity mapping loss, face feature loss, L1 term, and cycle-consistency loss leads to 5%, 12%, 13%, and 9% lower rank-1 identification rate than using our full method.

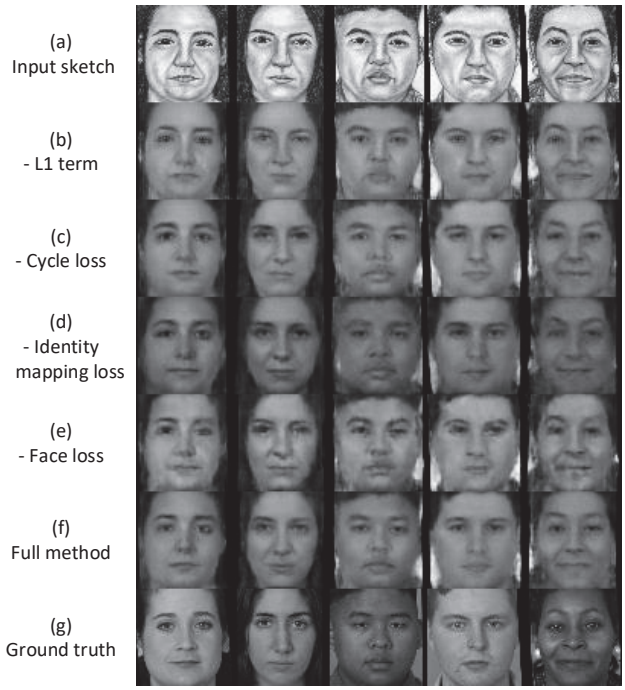


Fig. 6. Example face photos synthesized by different methods in the ablation study. (a) input face sketches from CUFSF, (b-e) synthesized face photos by the incomplete method without L1 distance term, cycle-consistency loss, identity mapping loss, face feature loss, respectively, (f) the generated photos by the proposed method, and (g) the ground-truth photos.

A careful observation of the synthesized sketches by different modes (some representative results can be seen in Fig. 6), shows that compared to other mehods, methods without face feature loss tend to produce evident noises and deformations. Our whole method generates the most visually pleasing photos among all these approaches. The CMC curves of individual methods are shown in Fig. 5. These results suggest the proposed approach with the designed feature-level loss is pertinent for the final recognition purpose. We have also performed the ablation study multiple times, the standard derivation of multiple tests is minor.

## V. CONCLUSION AND FUTURE WORK

While face sketch-photo synthesis has wide applications in law enforcement, forensics and entertainment, it remains challenging, due to the big modality gap between photos and sketches. In this work, we have proposed an adversarial model for synthesizing high-quality face photos from sketches. Besides the commonly used discriminators for evaluating the authenticity of the generators, we have

introduced a feature-level loss to assure higher similarities between the synthesized photos and the ground-truth photos. Experimental results on the public-domain database CUFSF show that the proposed approach outperforms the state-of-the-art methods in sketch-to-photo transformation w.r.t. SSIM. Face identification using the synthesized photos in a gallery set with 10K background images show our approach is instrumental in improving sketch based face recognition accuracy.

We notice that during the network training, the performance may vary within few epochs. A possible reason lies in the relatively small dataset size employed for sketch-to-photo transformation. Given the limited data, in our future work we intend to investigate different pre-training methods to allow for better convergence of the sketch-to-photo transformation model. In addition, we would like to explore the possibility of synthesizing photos from more challenging forensic sketches (i.e., the sketches drawn based on eyewitness's verbal descriptions, and composite sketches [42]), in which sketches can be very different from photos.

## ACKNOWLEDGEMENT

This research was supported in part by the Natural Science Foundation of China (grants 61732004 and 61672496), External Cooperation Program of Chinese Academy of Sciences (CAS) (grant GJHZ1843), Strategic Priority Research Program of CAS (grant XDB02070004), and Youth Innovation Promotion Association CAS (2018135).

## REFERENCES

[1] X. Tang and X. Wang, "Face sketch synthesis and recognition," in *IEEE ICCV*, pp. 687–694, 2003.
[2] X. Tang and X. Wang, "Face sketch recognition," *IEEE Trans. CSVT*, vol. 14, no. 1, pp. 50–57, 2004.
[3] Q. Liu, X. Tang, H. Jin, H. Lu, and S. Ma, "A nonlinear approach for face sketch synthesis and recognition," in *IEEE CVPR*, pp. 1005–1010, 2005.
[4] W. Liu, X. Tang, and J. Liu, "Bayesian tensor inference for sketch-based facial photo hallucination.," in *IJCAI*, pp. 2141–2146, 2007.
[5] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Trans. PAMI*, vol. 31, no. 11, pp. 1955–1967, 2009.
[6] W. Zhang, X. Wang, and X. Tang, "Coupled information-theoretic encoding for face photo-sketch recognition," in *IEEE CVPR*, pp. 513–520, 2011.
[7] B. F. Klare and A. K. Jain, "Heterogeneous face recognition using kernel prototype similarities," *IEEE Trans. PAMI*, vol. 35, no. 6, pp. 1410–1422, 2013.
[8] H. Han, B. F. Klare, K. Bonnen, and A. K. Jain, "Matching composite sketches to face photos: A component-based approach," *IEEE Trans. IFS*, vol. 8, no. 1, pp. 191–204, 2013.
[9] T. Chugh, M. Singh, S. Nagpal, R. Singh, and M. Vatsa, "Transfer learning based evolutionary algorithm for composite face sketch recognition," in *IEEE CVPR Workshops*, pp. 619–627, 2017.
[10] S. Nagpal, M. Singh, R. Singh, M. Vatsa, A. Noore, and A. Majumdar, "Face sketch matching via coupled deep transform learning," in *IEEE ICCV*, pp. 5429–5438, 2017.
[11] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE CVPR*, pp. 2223–2232, 2017.
[12] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.
[13] Y. Song, L. Bao, Q. Yang, and M.-H. Yang, "Real-time exemplar-based face sketch synthesis," in *ECCV*, pp. 800–813, 2014.
[14] L. Chang, M. Zhou, Y. Han, and X. Deng, "Face sketch synthesis via sparse representation," in *IEEE ICIP*, pp. 2146–2149, 2010.
[15] N. Wang, X. Gao, D. Tao, and X. Li, "Face sketch-photo synthesis under multi-dictionary sparse representation framework," in *IEEE ICIG*, pp. 82–87, 2011.
[16] S. Wang, L. Zhang, Y. Liang, and Q. Pan, "Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis," in *IEEE CVPR*, pp. 2216–2223, 2012.
[17] M. Zhu and N. Wang, "A simple and fast method for face sketch synthesis," in *ACM ICIMCS*, pp. 168–171, 2016.
[18] L. Zhang, L. Lin, X. Wu, S. Ding, and L. Zhang, "End-to-end photo-sketch generation via fully convolutional representation learning," in *ACM ICMR*, pp. 627–634, 2015.
[19] M. Zhu, N. Wang, X. Gao, and J. Li, "Deep graphical feature learning for face sketch synthesis," in *IJCAI*, pp. 3574–3580, 2017.
[20] I. Jolliffe, *Principal component analysis*. Springer, 2002.
[21] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
[22] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *NIPS*, pp. 801–808, 2007.
[23] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. IP*, vol. 15, no. 12, pp. 3736–3745, 2006.
[24] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *IEEE Trans. IP*, vol. 17, no. 1, pp. 53–69, 2008.
[25] W. Dong, L. Zhang, G. Shi, and X. Wu, "Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization," *IEEE Trans. IP*, vol. 20, no. 7, pp. 1838–1857, 2011.
[26] W. Dong, L. Zhang, and G. Shi, "Centralized sparse representation for image restoration," in *IEEE ICCV*, pp. 1259–1266, 2011.
[27] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, pp. 2672–2680, 2014.
[28] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *IEEE ICCV*, pp. 2813–2821, 2017.
[29] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *IEEE CVPR*, pp. 1125–1134, 2017.
[30] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *MICCAI*, pp. 234–241, 2015.
[31] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *IEEE CVPR*, pp. 2414–2423, 2016.
[32] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *ECCV*, pp. 694–711, 2016.
[33] X. Wu, R. He, and Z. Sun, "A lightened cnn for deep face representation," *Computer Science*, 2015.
[34] S. J. Klum, H. Han, B. F. Klare, and A. K. Jain, "The FaceSketchID system: Matching facial composites to mugshots," *IEEE Trans. TIFS*, vol. 9, no. 12, pp. 2248–2263, 2014.
[35] H. S. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa, "Memetically optimized mcwld for matching sketches with digital face images," *IEEE Trans. IFS*, vol. 7, no. 5, pp. 1522–1535, 2012.
[36] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep. 07-49, University of Massachusetts, Amherst, October 2007.
[37] K. Ricanek and T. Tesafaye, "Morph: A longitudinal image database of normal adult age-progression," in *IEEE FG*, pp. 341–345, 2006.
[38] H. Kazemi, M. Iranmanesh, A. Dabouei, S. Soleymani, and N. M. Nasrabadi, "Facial attributes guided deep sketch-to-photo synthesis," in *IEEE WACV Workshops*, pp. 1–8, 2018.
[39] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *Trans. IP*, vol. 13, no. 4, pp. 600–612, 2004.
[40] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *NIPS*, pp. 2234–2242, 2016.
[41] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *IEEE CVPR*, pp. 4681–4690, 2017.
[42] P. Mittal, A. Jain, G. Goswami, R. Singh, and M. Vatsa, "Recognizing composite sketches with digital face images via ssd dictionary," in *IEEE IJCB*, pp. 1–6, 2014.