2018 IEEE Winter Conference on Applications of Computer Vision Workshops

Facial Attributes Guided Deep Sketch-to-Photo Synthesis

Hadi Kazemi Mehdi Iranmanesh Ali Dabouei Sobhan Soleymani Nasser M. Nasrabadi
West Virginia University
{hakazemi, seiranmanesh, ad0046, ssoleyma}@mix.wvu.edu, nasser.nasrabadi@mail.wvu.edu

Abstract

Face sketch-photo synthesis is a critical application in law enforcement and digital entertainment industry. Despite the significant improvements in sketch-to-photo synthesis techniques, existing methods have still serious limitations in practice, such as the need for paired data in the training phase or having no control on enforcing facial attributes over the synthesized image. In this work, we present a new framework, which is a conditional version of CycleGAN, conditioned on facial attributes. The proposed network forces facial attributes, such as skin and hair color, on the synthesized photo and does not need a set of aligned face-sketch pairs during its training. We evaluate the proposed network by training on two real and synthetic sketch datasets. The hand-sketch images of the FERET dataset and the color face images from the WVU Multi-modal dataset are used as an unpaired input to the proposed conditional CycleGAN with the skin color as the controlled face attribute. For more attribute guided evaluation, a synthetic sketch dataset is created from the CelebA dataset and used to evaluate the performance of the network by forcing several desired facial attributes on the synthesized faces.

1. Introduction

Automatic face sketch-photo or photo-sketch synthesis and identification have always been important topics in computer vision and machine learning due to their vital applications in law enforcement and digital entertainment industry [31, 16, 20, 27]. In law enforcement, in most cases, the photo of a suspect is not available in the police database and therefore, the forensic sketch, which is drawn by a police artist based on an eyewitness testimony, is the only clue to identify the suspect. However, recognition of the suspect using a face sketch is much harder than a face photo because of the significant differences between the two modalities, such as the texture and geometric mismatching, which reduces the chance of identifying the suspect by person or from a mugshot database. In these cases, automatic face sketch-photo synthesis comes handy by generating suspects' photos from their forensic sketches.

Most of the previous research works on the sketch-based photo synthesis have addressed the problem of sketch-photo synthesis by using pairs of sketches and photos which are captured under highly controlled conditions, *i.e*., neutral expression and frontal pose. Different techniques have been studied including sparse representations [3], transductive learning of a probabilistic sketch-photo generation model [17], support vector regression [37], Bayesian tensor inference [32], embedded hidden Markov model [30], multiscale Markov random field model [34]. However, slight changes in the conditions can drastically degrade the performance of these photo synthesizing methods which are based on the existence of such controlled training pairs.

Recently, several studies proposed new methods which are more robust to the variation in different conditions [15, 40]. In [7], a deep convolutional neural network (DCNN) is utilized to tackle the problem of face sketch-photo synthesis in uncontrolled conditions. They developed three different models for multiple sketch styles. Peng *et al*. [22] derived a high dimensional multi-view feature vector using multiple filters and multiple local features to reduce the influence of different lighting conditions. The Locality Preserving Projections (LPP) is then adopted to reduce the feature dimensionality. Finally, traditional sketch synthesis methods such as Markov Random Fields (MRF) and Markov Weight Fields (MWF) are utilized to generate a photo from the multi-view feature representation. Similarly, [21] used multiple representations in an MRF model to gain robustness to lighting and pose variations.

DCNNs have been successfully applied in many cross-modality tasks [26, 10], specially in image transformation tasks such as sketch-photo and photo-sketch synthesis. The key to their success is in their ability of modeling nonlinear spacial transform between their input and output domains [2]. Zhang *et al*. [38] used a six-layer convolutional neural network (CNN) to generate sketches from photos. In [38], a new optimization objective function is utilized in the form of joint generative discriminative minimization to preserve the person's identity. A CNN-based framework was presented in [4] to transfer image style between arbitrary

978-1-5386-5188-9/18/$31.00 ©2018 IEEE
DOI 10.1109/WACVW.2018.00006

images by learning generic feature representations. Furthermore, a combination of a deep neural network with classical MRFs based texture synthesis was used in [14] to transfer the image style between arbitrary images or sketches.

For many years, the main objective function of CNNs in image generation applications has been defined as minimization of the Euclidean distance between the predicted and ground truth pixels. However, a network that is trained based on this objective function tends to generate blurry images [39]. More recently, deep convolutional generative adversarial networks (GANs) [5] have led to a significant improvement in image generation tasks by selecting a new loss function to generate more sharp and realistic images. This is done by attempting to fool a discriminator network that distinguishes between synthetic images and the real ones. GANs initially were developed in an unconditional setting to learn the distribution of the training data [5]. Fortunately, conditional GANs (cGAN) [11] are also introduced in the literature that learn conditional generative models and generate images conditioned on an input. This makes cGANs a good fit for many image transformation applications such as sketch-photo synthesis [25], image inpainting [36], general-purpose image-to-image translation [11], image manipulation [43], and style transfer [29]. Among them, Sangkloy *et al*. [25] specifically studied the application of cGANs on face sketch-photo synthesis.

Despite the success of the aforementioned techniques on image synthesis applications, they mostly suffer from a major drawback: They need the corresponding pair of images from both the source and the target modalities to train the network. Unfortunately, this is difficult to meet in practice since each artist has its own painting style, and the trained networks usually need to be fine tuned using scarcity-based domain adaptation techniques [19] for a new unseen sketch style. Besides, there are usually many suspects' sketches without their ground truth photos as they have not been caught yet. In order to solve the problem of paired training data unavailability, an unpaired image-to-image translation framework was proposed in [44], so called CycleGAN. They proposed an approach to learn image translation from a source domain to a target domain without any paired examples. For the same reason, in this paper, we follow the same approach as CycleGAN to train a network for sketch-photo synthesis in absence of paired samples.

Given the impressive results of recent face sketch-photo synthesis works, there is still a missing key part in this process which is conditioning the face synthesis task on the soft biometric traits. Especially in the application of sketch-photo synthesis, based on the quality of sketches, there are usually some face attributes which are missing in the painted sketches, such as skin, hair, eye colors, gender, and ethnicity. Furthermore, conditioning the image synthesis process to other adhered facial characteristics, such as having eyeglasses or a hat, provide extra information about the individual of interest and can result in a more precise and higher quality synthesized output. Consequently, describing and manipulating attributes from face images have been active research topics for years [41, 12, 28]. The application of soft biometric traits in person identification has also been studied in the literature [42]. Face attributes help to construct face representations and train domain classifiers for identity prediction. However, few researchers have addressed this problem in sketch-photo synthesis [8], attribute-image synthesis [35], and face editing [23, 13]. Despite this interest, no one to the best of our knowledge has proposed an unpaired sketch-photo synthesis scheme disentangled with respect to relevant facial attributes.

Although the CycleGAN solved the problem of learning a GAN network in the absence of paired training data, the original version does not force any conditions, *e.g*., facial attributes, on the image synthesis process. In this paper, we propose a new framework built on the CycleGAN to generate face photos from sketches conditioned on relevant facial attributes. To this end, we developed a conditional version of the CycleGAN which we refer to as the cCycleGAN and trained it by an extra discriminator to force the desired facial attributes on the synthesized images. The main contributions of this paper include the following:

- We propose a novel framework for facial attribute guided Sketch-Photo synthesis.
- We introduce a new version of CycleGAN with conditional setting. Adding conditions to the CycleGAN improves the stability of the network suring its training phase as the missing information in any of the source or target domains could make the training process unstable.
- The proposed attribute learning framework does not need a paired training data which allow us to train the network even on sketch datasets without their corresponding ground truth photos.

2. Conditional Generative Adversarial Networks (cGANs)

GANs [5] are a group of generative models which learn to map a random noise z to output image y: $G(z) : z \longrightarrow y$. They can be extended to a conditional GAN (cGAN) if the generator model, G, (and usually the discriminator) is conditioned on some extra information, x, such as an image or class labels. In other words, cGAN learns a mapping from an input x and a random noise z to the output image y: $G(x, z) : \{x, z\} \longrightarrow y$. The generator model is trained to generate an image which is not distinguishable from "real" samples by a discriminator network, D. The

discriminator is trained adversarially to discriminate between the "fake" generated images by the generator and the real samples from the training dataset. Both the generator and the discriminator are trained simultaneously following a two-player min-max game.

The objective function of cGAN is defined as:

$$l_{GAN}(G,D) = \mathbf{E}_{x,y\sim p_{data}}[\log D(x,y)] + \mathbf{E}_{x,z\sim p_z}[\log(1-D(x,G(x,z)))], \quad (1)$$

where G attempts to minimize it and D tries to maximize it. Previous works in the literature have found it beneficial to add an extra $L2$ or $L1$ distance term to the objective function which forces the network to generate images which are near the ground truth. Isola *et al.* [11] found $L1$ to be a better candidate as it encourages less blurring in the generated output. In summary, the generator model is trained as follows:

$$G^* = \arg\min_G \max_D l_{GAN}(G,D) + \lambda l_{L1}(G), \quad (2)$$

where λ is a weighting factor and $l_{L1}(G)$ is

$$l_{L1}(G) = \| \, y - G(x,z) \, \|_1 \, . \quad (3)$$

2.1. Training Procedure

In each training step, an input, x is passed to the generator to produce the corresponding output, $G(x,z)$. The generated output and the input are concatenated and fed to the discriminator. First, the discriminator's weight is updated in a way to distinguish between the generated output and a real sample from the target domain. Then, the generator is trained to fool the discriminator by generating more realistic images.

3. CycleGAN

The main goal of CycleGAN [44] is to train two generative models, G_x and G_y. These two models learn the mapping functions between two domains x and y. The model, as illustrated in Figure 1, includes two generators; the first one maps x to y: $G_y(x) : x \longrightarrow y$ and the other does the inverse mapping y to x: $G_x(y) : y \longrightarrow x$. There are two adversarial discriminators D_x and D_y, one for each generator. More precisely, D_x distinguishes between "real" x samples and its generated "fake" samples $G_x(y)$, and similarly, D_y discriminates between "real" y and the "fake" $G_y(x)$. Therefore, there is a distinct adversarial loss in CycleGAN for each of the two (G_x, D_x) and (G_y, D_y) pairs. Notice that the adversarial losses are defined as in Eq. 1.

For a high capacity network to be trained using only the adversarial loss, there is a possibility of mapping the same set of inputs to a random permutation of images in the target domain. In other words, the adversarial loss is not enough

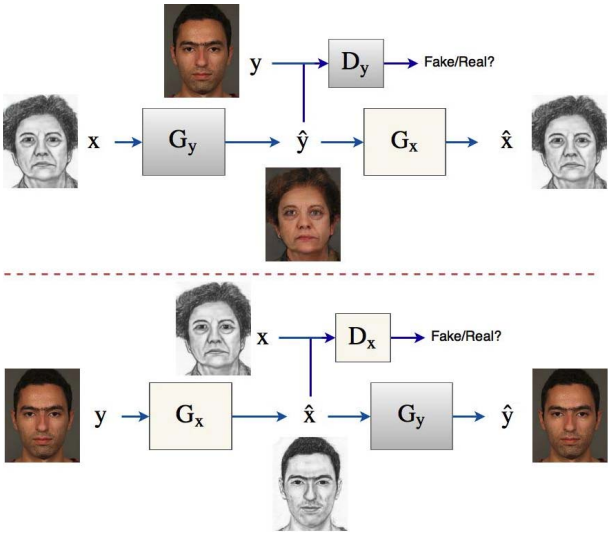

Figure 1. CycleGAN

to guarantee that the trained network generates the desired output. This is the reason behind having an extra $L1$ distance term in the objective function of cGAN as shown in Eq. 2. As shown in Figure 1, in the case of CycleGAN, there are no paired images between the source and target domains, which is the main feature of CycleGAN over cGAN. Consequently, the $L1$ distance loss cannot be applied to this problem. To tackle this issue, a cycle consistency loss was proposed in [44] which forced the learned mapping functions to be cycle-consistent. Particularly, the following conditions should be satisfied

$$\begin{aligned} x &\longrightarrow G_y(x) \longrightarrow G_x(G_y(x)) \approx x \\ y &\longrightarrow G_x(y) \longrightarrow G_y(G_x(y)) \approx y. \end{aligned} \quad (4)$$

To this end, a cycle consistency loss is defined as

$$\begin{aligned} l_{cyc}(G_x, G_y) = \mathbf{E}_{x\sim p_{data}}[\|x - G_x(G_y(x))\|_1] \\ + \mathbf{E}_{y\sim p_{data}}[\|y - G_y(G_x(y))\|_1]. \end{aligned} \quad (5)$$

Taken together, the full objective function is

$$\begin{aligned} l_(G_x, G_y, D_x, D_y) = l_{GAN}(G_x, D_x) + l_{GAN}(G_y, D_y) \\ + \lambda l_{cyc}(G_x, G_y), \end{aligned} \quad (6)$$

where λ is a weighting factor to control the importance of the objectives and the whole model is trained as follows

$$G_x^*, G_y^* = \arg\min_{G_x,G_y} \max_{D_x,D_y} l(G_x, G_y, D_x, D_y). \quad (7)$$

From now on, we use x for our source domain which is the sketch domain and y for the target domain or the photo domain.

Layer #	Type	Kernel	Input Size	Output Size
1	Conv	7x7	128 x 128	128 x 128
2	Conv	3x3	128 x 128	64 x 64
3	Conv	3x3	64 x 64	32 x 32
4-12	Res	5x5	32 x 32	32 x 32
13	Conv	3x3	32 x 32	64 x 64
14	Conv	3x3	64 x 64	128 x 128
15	Conv	7x7	128 x 128	128 x 128

Table 1. Generator architecture with an input of size 128x128

Layer #	Type	Kernel	Input Size	Output Size
1	Conv	4x4	128 x 128	64 x 64
2	Conv	4x4	64 x 64	32 x 32
3	Conv	4x4	32 x 32	16 x 16
4	Conv	4x4	16 x 16	15 x 15
5	Conv	4x4	15 x 15	14 x 14

Table 2. Discriminator architecture with an input of size 128x128

3.1. Architecture

The two generators, G_x and G_y, adopt the same architecture [44] consisting of six convolutional layers and nine residual blocks [9]. Table 1 details the generators' architecture. The discriminators also share their architecture which is summarized in Table 2. The output of the discriminator is of size 30x30. Each output pixel corresponds to a patch of the input image and tries to classify if the patch is real or fake. More details are reported in [44].

4. Conditional CycleGAN (cCycleGAN)

The CycleGAN architecture has solved the problem of having unpaired training data, but still, has a major drawback: Extra conditions, such as soft biometric traits, cannot be forced on the target domain. To tackle this problem, we proposed a CycleGAN architecture with a soft biometrics conditional setting which we refer it as Conditional CycleGAN (cCycleGAN). Since in the sketch-photo synthesis problem, attributes (*e.g.*, skin color) are missing on the sketch side and not on the photo side, the photo-sketch generator, $G_x(y)$, is left unchanged in the new setting. However, the sketch-photo generator, $G_y(x)$, needs to be modified by conditioning it on the facial attributes. The new sketch-photo generator maps (x, a) to y, *i.e.*, $G_y(x, a) : (x, a) \longrightarrow y$, where a stands for the desired facial attributes to be present in the synthesized photo. The corresponding discriminator, $D_y(x, a)$, is also conditioned on both the sketch, x, and the desired facial attributes, a. The definition of the loss function remains the same as in CycleGAN given by Eq. 6.

Despite the previous work in face editing [23], our preliminary results showed that having only a single discriminator conditioned on the desired facial attributes was not enough to force the attributes on the generator's output of the CycleGAN. Consequently, instead of increasing the complexity of the discriminator, we trained an additional auxiliary discriminator, $D_a(y, a)$, to detect if the desired attributes are present in the synthesized photo or not. In other words, the sketch-photo generator, $G_y(x, a)$, tries to fool an extra attribute discriminator, $D_a(y, a)$, which checks the presence of the desired facial attributes. The objective function of the attribute discriminator is defined as follows:

$$\begin{aligned} l_{Att}(G_y, D_a) = &\mathbf{E}_{a,y\sim p_{data}}[\log D_a(a, y)] + \\ &\mathbf{E}_{y\sim p_{data}, \bar{a}\neq a}[\log(1 - D_a(\bar{a}, y))] + \\ &\mathbf{E}_{a,y\sim p_{data}}[\log(1 - D_a(a, G_y(x, a)))], \end{aligned} \tag{8}$$

where a is the corresponding attributes of the real image, y, and $\bar{a} \neq a$ is a set of random arbitrary attributes. Therefore, the total loss of the cCycleGAN is

$$\begin{aligned} l_(G_x, G_y, D_x, D_y) = l_{GAN}(G_x, D_x) + l_{GAN}(G_y, D_y) \\ + \lambda_1 l_{cyc}(G_x, G_y) + \lambda_2 l_{Att}(G_y, D_a), \end{aligned} \tag{9}$$

where λ_1 and λ_2 are weighting factors to control the importance of the objectives.

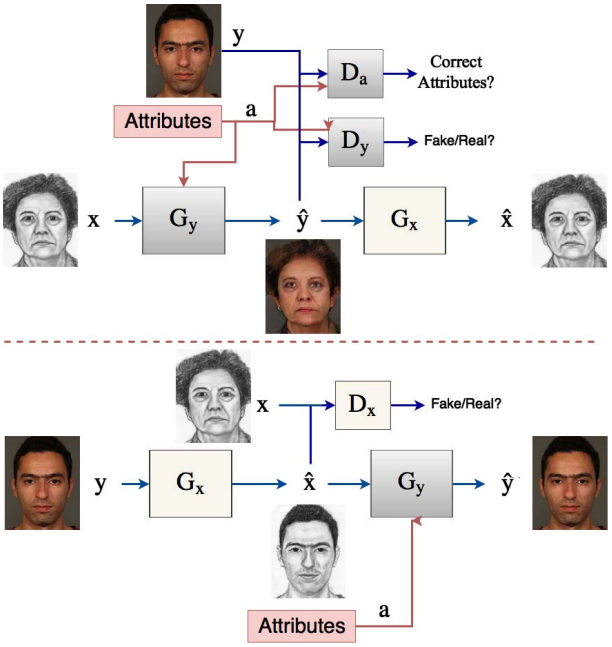

Figure 2. cCycleGAN architecture, including Sketch-Photo cycle (top) and Photo-Sketch cycle (bottom).

4.1. Architecture

Our proposed cCycleGAN adopts the same architecture as in CycleGAN. However, to condition the generator and

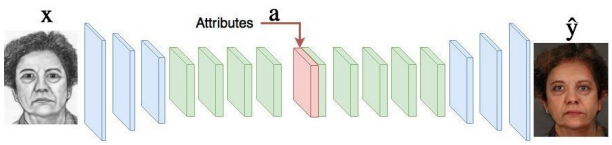

Figure 3. Sketch-Photo generator network, $G_y(x, a)$, in cCycleGAN.

Figure 4. Samples from FERET, WVU Multi-modal, CelebA, and X-DOG generated synthetic sketches datasets.

the discriminator to the facial attributes, we slightly modified the architecture. The generator which transforms photos into sketches, $G_x(y)$, and its corresponding discriminator, D_x, are left unchanged as there is no attribute to force in sketch generation phase. However, in the sketch-photo generator, $G_y(x)$, we insert the desired attributes before the fifth residual block of the bottleneck (Figure 2). To this end, each attribute is repeated 4096 (64*64) times and then resized to a matrix of size 64×64. Then all of these attribute feature maps and the output feature maps of the fourth residual block are concatenated in depth and passed to the next block, as shown in Figure 3. The same modification is applied to the corresponding attribute discriminator, D_a. All the attributes are repeated, resized, and concatenated with the generated photo in depth and are passed to the discriminator.

4.2. Training Procedure

We follow the same training procedure as in Section 2.1 for the photo-sketch generator. However, for the sketch-photo generator, we need a different training mechanism to force the desired facial attributes to be present in the generated photo. Therefore, we define a new type of negative sample for the attribute discriminator, D_a, which is defined as a real photo from the target domain but with a wrong set of attributes, $\bar{a}$. The training mechanism forces the sketch-photo generator to produce faces with the desired attributes. At each training step, this generator synthesizes a photo with the same attributes, a, as the real photo. Both the corresponding sketch-photo discriminator, D_y, and attribute discriminator, D_a, are supposed to detect the synthesized photo as a fake sample. The attribute discriminator, D_a, is also trained with two other pairs: a real photo with correct attributes as a real sample, and a real photo with wrong set of attributes as a fake sample. Simultaneously, the sketch-photo generator attempts to fool both of the discriminators.

5. Experimental Results

5.1. Datasets

FERET Sketch: The FERET database [24] includes 1,194 sketch-photo pairs. Sketches are hand-drawn by an artist while looking at the face photos. Both the face photos and sketches are grayscale images of size 250 × 200 pixels. However, to produce color photos we did not use the grayscale face photos of this dataset to train the cCycleGAN. We randomly selected 1000 sketches to train the network and the remaining 194 are used for testing.

WVU Multi-modal: To synthesis color images from the FERET sketches, we use the frontal view face images from WVU Multi-modal [1]. The Dataset contains 3453 high-resolution color frontal images of 1200 subjects. The images are aligned, cropped and resized to the same size as FERET Sketch, *i.e.*, 250 × 200 pixels. The dataset does not contain any facial attributes. However, for each image, the average color of a 25 × 25 pixels rectangular patch (placed in forehead or cheek) is considered as the skin color. Then, they are clustered into three classes, namely white, brown and black, based on their intensities.

CelebFaces Attributes (CelebA): We use the aligned and cropped version of the CelebA dataset [18] and scale the images down to 128 × 128 pixels. We also randomly split it into two partitions, 182K for training and 20K for testing. Of the original 40 attributes, we selected only those attributes that have a clear visual impact on the synthesized faces and are missing in the sketch modality, which leaves a total of six attributes, namely black hair, brown hair, blond hair, gray hair, pale skin, and gender. Due to the huge differences in face views and the background in FERET and celebA databases, the preliminary results did not show an acceptable performance on FERET-celebA pair training. Consequently, we generated a synthetic sketch dataset by applying xDOG [33] filter to the celebA dataset. However, to train the cCycleGAN, the synthetic sketch and photo images are used in an unpaired fashion. Figure 4 illustrates a sample from each of the datasets.

5.2. Results on FERET and WVU Multi-modal

Sketches from the FERET dataset are trained in couple with frontal face images from the WVU Multi-modal to train the proposed cCycleGAN. Since there is no facial attributes associated with the color images of the WVU Multi-modal dataset, we have classified them based on their skin colors. Consequently, the skin color is the only attribute which we can control during the sketch-photo synthesis. Therefore, the input to the sketch-photo generator has two channels including a gray-scale sketch image, x, and a sin-

gle attribute channel, a, for the skin color. The sketch images are normalized to stand in $[-1, 1]$ range. Similarly, the skin color attribute gets -1, 0, and 1 for the black, brown and white skin colors, respectively. Figure 5 shows the results of the cCycleGAN after 200 epochs on the test data. The three skin color classes are not represented equally in the dataset which obviously balanced the results towards the lighter skins.

5.3. Results on CelebA and synthesized sketches

Preliminary results reveal that the CycleGAN training can get unstable when there is a significant difference, such as differences in scale and face poses, in the source and target datasets. The easy task of the discriminator in differentiating between the synthesized and real photos in these cases could account for this instability. Consequently, we generated a synthetic sketch dataset as a replacement to the FERET dataset. Among the 40 attributes provided in the CelebA dataset, we have selected the six most relevant ones in terms of the visual impacts on the sketch-photo synthesis, including black hair, blond hair, brown hair, gray hair, male, and pale skin. Therefore, the input to the sketch-photo generator has seven channels including a gray-scale sketch image, x, and six attribute channels, a. The attributes in CelebA dataset are binary, we have chosen -1 for a missing attribute and 1 for an attribute which is supposed to be present in the synthesized photo. Figure 6 shows the results of the cCycleGAN after 50 epochs on the test data. The trained network can follow the desired attributes and force them on the synthesized photo.

5.4. Evaluation of synthesized photos with a face verifier

For the sake of evaluation, we utilized a VGG16-based face verifier pre-trained on the CMU Multi-PIE dataset [6]. To evaluate the proposed algorithm, we first selected the identities which had more than one photos in the testing set. Then, for each identity, one photo is randomly added to the test gallery, and a synthetic sketch corresponding to another photo of the same identity is added to the test prob. Finally, every prob synthetic sketch is given to our attribute-guided sketch-photo synthesizer and the resulting synthesized photos are used for face verification against the entire test gallery. This evaluation process was repeated 10 times. Table 3 depicts the face verification accuracies of the proposed attribute-guided approach and the results of the original cycle-GAN on celebA dataset. The results of our proposed network significantly improved on the original cycle-GAN with no attribute information.

Table 3. Verification performance of the proposed ccycle-GAN network vs. the original cycle-GAN

Method	Accuracy (%)
cycle-GAN	%61.34 $\pm$ 1.05
ccycle-GAN	%65.53 $\pm$ 0.93

6. Conclusion

In this paper, we presented a conditional CycleGAN (cCycleGAN) for soft biometrics (facial attributes) guided unpaired face sketch-photo synthesis problem. To this end, an additional auxiliary attribute discriminator was utilized with an appropriate loss to force the desired facial attributes on the output of the generator. The pair of real face photo from the training data with a set of false attributes defined a new fake input to the attribute discriminator in addition to the pair of generator's output and a set of random attributes. The proposed network was trained on two pairs of hand-drawn and synthetic sketch datasets in an unpaired fashion. Our experiments reveal how the network can generate multiple photos with quite different facial attributes per single sketch. However, similar to previous proposed solution for image editing, editing an attribute can cause unwanted structural edition of the image in some areas. Going forward, we would like to force the network to keep the exact face style in the output while editing the desired attributes on the synthesized photo.

References

[1] Biometrics and identification innovation center, wvu multimodal dataset. Available at `http://biic.wvu.edu/`,.

[2] A. Dabouei, H. Kazemi, M. Iranmanesh, and N. M. Nasrabadi. Fingerprint distortion rectification using deep convolutional neural networks. In *Biometrics (ICB), 2018 International Conference on*. IEEE, 2018.

[3] X. Gao, N. Wang, D. Tao, and X. Li. Face sketch–photo synthesis and retrieval using sparse representation. *IEEE Transactions on circuits and systems for video technology*, 22(8):1213–1226, 2012.

[4] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[6] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.

[7] Y. Güçlütürk, U. Güçlü, R. van Lier, and M. A. van Gerven. Convolutional sketch inversion. In *European Conference on Computer Vision*, pages 810–824. Springer, 2016.

[8] Q. Guo, C. Zhu, Z. Xia, Z. Wang, and Y. Liu. Attribute-controlled face photo synthesis from simple line drawing. *arXiv preprint arXiv:1702.02805*, 2017.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE con-*

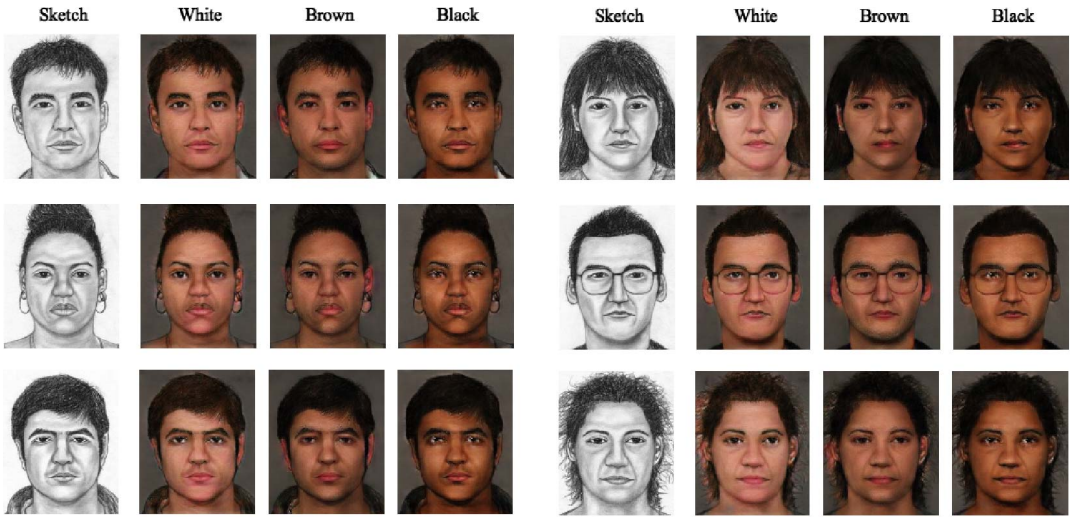

Figure 5. Sketch-based photo synthesis of hand-drawn test sketches from FERET dataset. Our network can adapt the synthesis results to satisfy different skin colors (white, brown, black).

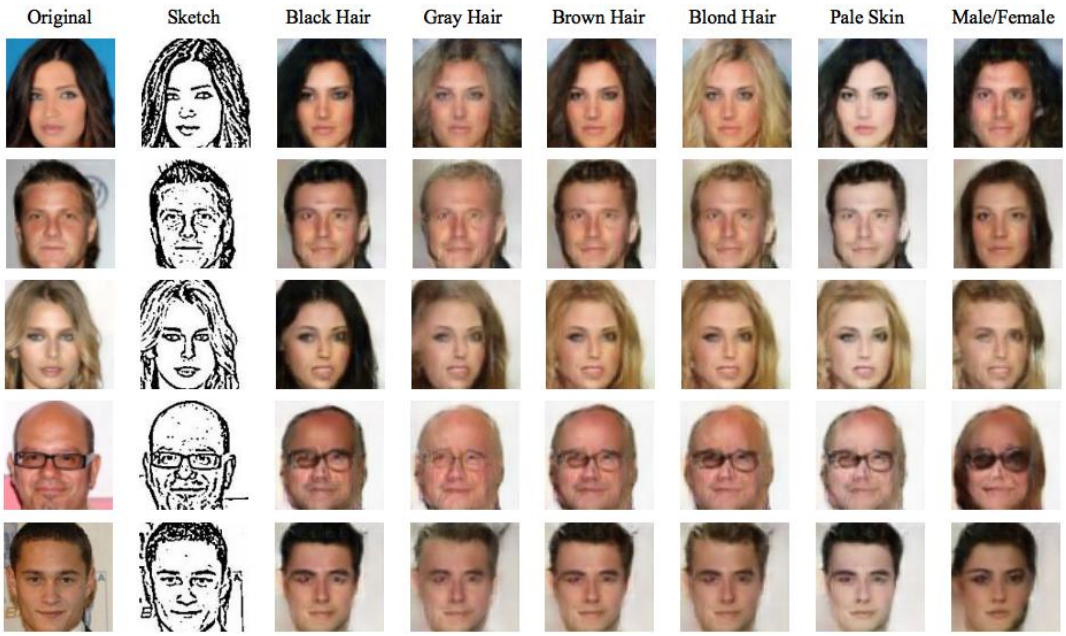

Figure 6. Attribute guided Sketch-based photo synthesis of synthetic test sketches from CelebA dataset. Our network can adapt the synthesis results to satisfy the desired attributes.

ference on computer vision and pattern recognition, pages 770–778, 2016.

[10] M. Iranmanesh, A. Dabouei, H. Kazemi, and N. M. Nasrabadi. Deep cross polarimetric thermal-to-visible face recognition. In *Biometrics (ICB), 2018 International Conference on*. IEEE, 2018.

[11] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016.

[12] A. Jourabloo, X. Yin, and X. Liu. Attribute preserved face de-identification. In *Biometrics (ICB), 2015 International Conference on*, pages 278–285. IEEE, 2015.

[13] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. Denoyer, and M. Ranzato. Fader networks: Manipulating images by sliding attributes. *arXiv preprint arXiv:1706.00409*, 2017.

[14] C. Li and M. Wand. Combining markov random fields and convolutional neural networks for image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2479–2486, 2016.

[15] Y.-h. Li, M. Savvides, and V. Bhagavatula. Illumination tolerant face recognition using a novel face from sketch synthesis approach and advanced correlation filters. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 2, pages II–II. IEEE, 2006.

[16] Q. Liu, X. Tang, H. Jin, H. Lu, and S. Ma. A nonlinear approach for face sketch synthesis and recognition. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 1005–1010. IEEE, 2005.

[17] W. Liu, X. Tang, and J. Liu. Bayesian tensor inference for sketch-based facial photo hallucination. pages 2141–2146, 2007.

[18] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

[19] S. Motiian, Q. Jones, S. Iranmanesh, and G. Doretto. Few-shot adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pages 6673–6683, 2017.

[20] C. Peng, X. Gao, N. Wang, and J. Li. Superpixel-based face sketch–photo synthesis. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(2):288–299, 2017.

[21] C. Peng, X. Gao, N. Wang, D. Tao, X. Li, and J. Li. Multiple representations-based face sketch–photo synthesis. *IEEE transactions on neural networks and learning systems*, 27(11):2201–2215, 2016.

[22] C. Peng, J. Li, N. Wang, and X. Gao. Multi-view representation based face sketch synthesis. In *Proceedings of International Conference on Internet Multimedia Computing and Service*, page 307. ACM, 2014.

[23] G. Perarnau, J. van de Weijer, B. Raducanu, and J. M. Álvarez. Invertible conditional GANs for image editing. *arXiv preprint arXiv:1611.06355*, 2016.

[24] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Transactions on pattern analysis and machine intelligence*, 22(10):1090–1104, 2000.

[25] P. Sangkloy, J. Lu, C. Fang, F. Yu, and J. Hays. Scribbler: Controlling deep image synthesis with sketch and color. *arXiv preprint arXiv:1612.00835*, 2016.

[26] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, pages 935–943, 2013.

[27] M. Song, C. Chen, J. Bu, and T. Sha. Image-based facial sketch-to-photo synthesis via online coupled dictionary learning. *Information Sciences*, 193:233–246, 2012.

[28] R. Tokola, A. Mikkilineni, and C. Boehnen. 3D face analysis for demographic biometrics. In *Biometrics (ICB), 2015 International Conference on*, pages 201–207. IEEE, 2015.

[29] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. S. Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. In *ICML*, pages 1349–1357, 2016.

[30] N. Wang, D. Tao, X. Gao, X. Li, and J. Li. Transductive face sketch-photo synthesis. *IEEE transactions on neural networks and learning systems*, 24(9):1364–1376, 2013.

[31] N. Wang, D. Tao, X. Gao, X. Li, and J. Li. A comprehensive survey to face hallucination. *International journal of computer vision*, 106(1):9–30, 2014.

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

[33] H. WinnemöLler, J. E. Kyprianidis, and S. C. Olsen. Xdog: an extended difference-of-gaussians compendium including advanced image stylization. *Computers & Graphics*, 36(6):740–753, 2012.

[34] B. Xiao, X. Gao, D. Tao, and X. Li. A new approach for face recognition by sketches in photos. *Signal Processing*, 89(8):1576–1588, 2009.

[35] X. Yan, J. Yang, K. Sohn, and H. Lee. Attribute2image: Conditional image generation from visual attributes. In *European Conference on Computer Vision*, pages 776–791. Springer, 2016.

[36] R. Yeh, C. Chen, T. Y. Lim, M. Hasegawa-Johnson, and M. N. Do. Semantic image inpainting with perceptual and contextual losses. *arXiv preprint arXiv:1607.07539*, 2016.

[37] J. Zhang, N. Wang, X. Gao, D. Tao, and X. Li. Face sketch-photo synthesis based on support vector regression. In *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pages 1125–1128. IEEE, 2011.

[38] L. Zhang, L. Lin, X. Wu, S. Ding, and L. Zhang. End-to-end photo-sketch generation via fully convolutional representation learning. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 627–634. ACM, 2015.

[39] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. In *European Conference on Computer Vision*, pages 649–666. Springer, 2016.

[40] W. Zhang, X. Wang, and X. Tang. Lighting and pose robust face sketch synthesis. In *European Conference on Computer Vision*, pages 420–433. Springer, 2010.

[41] Y. Zhong, J. Sullivan, and H. Li. Face attribute prediction using off-the-shelf CNN features. In *Biometrics (ICB), 2016 International Conference on*, pages 1–7. IEEE, 2016.

[42] J. Zhu, S. Liao, D. Yi, Z. Lei, and S. Z. Li. Multi-label CNN based pedestrian attribute learning for soft biometrics. In *Biometrics (ICB), 2015 International Conference on*, pages 535–540. IEEE, 2015.

[43] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros. Generative visual manipulation on the natural image manifold. In *European Conference on Computer Vision*, pages 597–613. Springer, 2016.

[44] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017.