End-to-End Deep Sketch-to-Photo Matching Enforcing Realistic Photo Generation

Leonardo Capozzi[1,2][0000−0002−5162−7124], João Ribeiro Pinto[1,2][0000−0003−4956−5902], Jaime S. Cardoso[1,2][0000−0002−3760−2473], and Ana Rebelo[1][0000−0003−4776−6057]

[1] INESC TEC, Porto, Portugal
{leonardo.g.capozzi, joao.t.pinto, jaime.cardoso, arebelo}@inesctec.pt
[2] Faculdade de Engenharia da Universidade do Porto, Porto, Portugal

Abstract. The traditional task of locating suspects using forensic sketches posted on public spaces, news, and social media can be a difficult task. Recent methods that use computer vision to improve this process present limitations, as they either do not use end-to-end networks for sketch recognition in police databases (which generally improve performance) or/and do not offer a photo-realistic representation of the sketch that could be used as alternative if the automatic matching process fails. This paper proposes a method that combines these two properties, using a conditional generative adversarial network (cGAN) and a pre-trained face recognition network that are jointly optimised as an end-to-end model. While the model can identify a short list of potential suspects in a given database, the cGAN offers an intermediate realistic face representation to support an alternative manual matching process. Evaluation on sketch-photo pairs from the CUFS, CUFSF and CelebA databases reveal the proposed method outperforms the state-of-the-art in most tasks, and that forcing an intermediate photo-realistic representation only results in a small performance decrease.

Keywords: Digital Forensics · Sketches · Generation.

1 Introduction

Over the years, convolutional neural networks (CNN) have been very successful for several pattern recognition and computer vision tasks, including that of face recognition. Recent publications in this topic often report significantly improved accuracy in the matching process when using deep learning methodologies [11, 2, 18]. However, face recognition based on photos or video is obviously an easier task than face recognition based forensic sketches, since a sketch might not be the most accurate representation of an individual, as it was drawn based on descriptions from eye witnesses [14]. On the problem of sketch-to-face matching, several recent state-of-the-art methods have also used CNNs to match a sketch to the corresponding identity [4, 10, 6, 1]. However, several of these do not take full advantage of the potential of deep learning as they are not end-to-end deep

approaches. The inclusion of blocks manually tuned or separately optimised can cause dissonance between different processes and limit achievable performance. One of the examples of separate processes in sketch-to-face recognition is the prior transformation of a sketch to resemble a real face photo. Several literature approaches apply such transformations including, sophisticated adversarial methods based on CycleGANs [6] and cGANs [10, 1]. Having a photo-realistic representation of the suspect's face is a great advantage for manual identification, when the automatic matching process fails to deliver useful results. Nevertheless, the separate optimisation of these processes will, as emphasized, induce performance limitations. Hence, this work tackles these two important aspects of forensic sketch-to-photo matching, that have not yet been addressed in the literature. The first aspect is avoiding the combination of separate processes that are individually optimised, which often limits achievable performance. The second aspect is enforcing the end-to-end model to offer an intermediate representation that is photo-realistic, so the authorities have access to a realistic face rendering that will help manual identification of the suspect. To achieve this, we propose an end-to-end model composed of a cGAN and a matching CNN that are jointly optimised. When trained, the model receives a sketch and returns a template that can be used for matching using simple distance metrics. Although the approach is end-to-end, the training strategy induces the cGAN to generate intermediate latent representations that look realistic and are similar to the corresponding real photographs. Hence, we avoid the performance limits often linked to non-end-to-end approaches, while retaining the intermediate realistic images that could help an alternative manual identification process.

2 Proposed Methodology

The proposed method is an end-to-end model that, although integrated and optimizable as a whole, is composed of two main parts: a sketch-to-render generator and a matching network (see Fig. 1). The sketch-to-render generator will transform the input sketch into a face rendering that is photo-realistic and similar to the real face that corresponds to the sketch. The matching network will receive the realistic rendering and output a template that can be used for matching through a simple distance measure. In the next subsections, the architecture of both parts, the loss function, and the training process of the model are described in higher detail.

2.1 Network Architecture

Sketch-to-render generator The sketch-to-render process is performed by an image-to-image model that receives a sketch and transforms it into a realistic face rendering. For this model, we use the generator of a cGAN [9], that follows the typical structure of a U-Net, with an encoder, that reduces data resolution at each level, followed by a decoder, that processes data up to the original resolution (see Fig. 2.1). Skip connections enable the transmission of

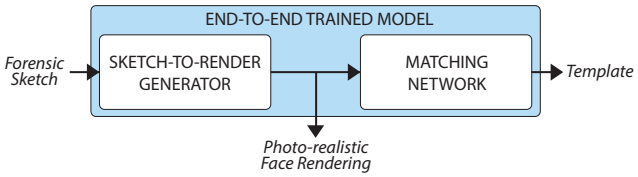

Fig. 1. Overview schema of the proposed method.

information between corresponding levels from the encoder to the decoder. The U-Net encoder mimicked the architecture of VGG-16 [17] and it is composed of 13 convolutional layers and 4 max-pooling layers distributed over 5 resolution levels. Convolutional layers have $64 - 512$ filters, with size of 3×3 and stride of 1×1. Max-pooling layers use window size and stride of 2×2. The decoder mirrors the structure of the encoder with convolutional and deconvolution layers. Convolutional layers have $64 - 512$ filters with size and stride similar to their respective encoder counterparts. Deconvolution layers have $64 - 512$ filters with size and stride of 2×2. The discriminator used during training is a CNN adapted from the cGAN of pix2pix [5] (see Fig. 2.1), which receives as input the photo-realistic representation outputted by the generator and the corresponding sketch, and outputs a prediction on whether it is real or generated. This discriminator is composed of 3 convolutional layers, 1 fully-connected layer, and batch normalization. The convolutional layers have $64 - 256$ filters with size of 4×4 and stride of 2×2.

Matching network After the sketch is transformed into a photo-realistic rendering of the face, the matching part of the model transforms this rendering into a template that can easily be used for matching. In this work, the matching network follows the structure of the VGG-16 [17] and uses pretrained VGG-Face weights [11]. Given a sketch, the output of the matching network (and of the method as a whole) is a template that can be used for matching. This template (or face descriptor) is a numerical uni-dimensional vector of 2622 features that describe the face represented on the input sketch. The sketch is thus matched with face photos from a database by computing the cosine distance between the respective templates.

2.2 Loss

The loss function used for training is a composition of several loss components relative to each part of the proposed model. The first component of the loss corresponds to the sketch-to-render, and comes from the typical training methodology of a cGAN. This part can be described as following:

$$\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{x,y}[\log D(x, y)] + \mathbb{E}_x[\log(1 - D(x, G(x)))], \quad (1)$$

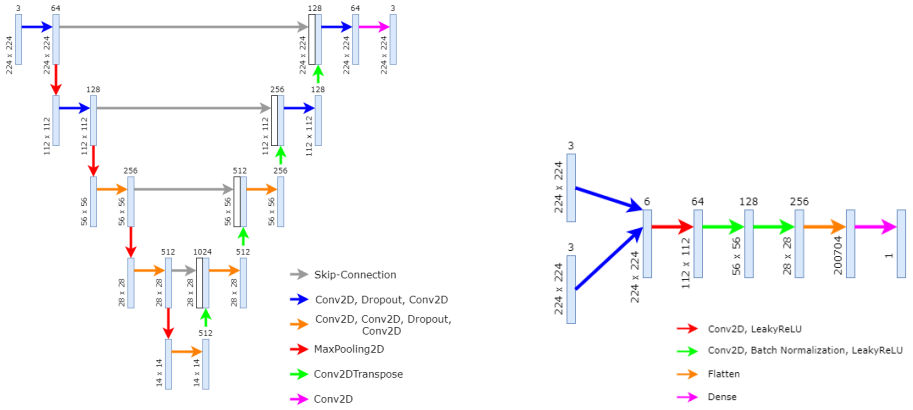

Fig. 2. Network architecture of the generator (on the left) and discriminator (on the right).

where x is a sketch and y is the corresponding ground-truth photo. The generator (G) tries to minimize this loss and the discriminator (D) competes to maximize it. This competition, if adequately balanced, will lead the generator to output increasingly realistic face images. Nevertheless, the generator should not only transform the sketch into a photo-realistic image, but also should closely mimic the respective ground-truth face photograph. In order to achieve this, a second loss term is defined as:

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y}[||y - G(x)||_1], \tag{2}$$

which minimizes the $L1$ norm of the difference between the real image (y) and the generated image ($G(x)$). Moreover, we need to ensure the identity information in the sketch is preserved and refined throughout the network, and matches that of the respective ground-truth image.

Hence, a third component is added to the loss function, corresponding to the matching part, such as:

$$\mathcal{L}_{match}(G) = \mathbb{E}_{x,y}[||V(y) - V(G(x))||_2]. \tag{3}$$

Notice that the weights of the VGG-Face network (V) are frozen, since we are using the pretrained weights. The only weights that are adjusted in the matching loss are the weights of the generator. Combining the loss components mentioned above, the final loss function becomes as:

$$\mathcal{L}(G, D) = \min_G \max_D \mathcal{L}_{cGAN}(G, D) + \lambda_1 \mathcal{L}_{L1}(G) + \lambda_2 \mathcal{L}_{match}(G). \tag{4}$$

3 Experimental Settings

3.1 Data

The proposed model has been trained using pairs of sketches and corresponding face images from the CUHK Face Sketch database (CUFS) [19], the CUHK Face Sketch FERET dataset (CUFSF) [19, 21] and the CelebA Dataset [8]. The CUFS dataset contains 606 sketch-photo pairs. The sketches correspond to face images acquired from students of the Chinese University of Hong Kong (CUHK) and from the AR and XM2GTS databases. Due to the current unavailability of the latter two databases, this work only used the 188 sketch-photo pairs relative to CUHK students. The original data split was used, which contained 88 pairs for model training and 100 pairs for testing. The CUFSF contains 1189 sketches that correspond to photos in the FERET database [13, 12]. Of those, 1089 pairs were used for training and 100 pairs were used for testing. The CelebA dataset is a large scale dataset, which contains 202599 face images of 10177 identities. We randomly select 2000 images for testing and the remaining for training. The selected images for testing contain 400 different identities and 5 images per identity. This dataset does not contain sketches, therefore we generated our own sketches using an edge detection algorithm [10].

3.2 Pre-processing

To uniformize the CUFS and CUFSF databases, the DeOldify API[3] was used to colorize the photos of CUFSF from the FERET database. The use of this API allows the proposed model to generate color images. After visually inspecting the colorized images, we noticed that the colorization process did not negatively affect their quality.

In style transfer problems using conditional GANs, image spatial consistency is paramount. When the locations of image landmarks are consistent between the input and the expected output, the network is able to offer realistic results. However, in the case of sketch-to-photo transformation, the shape, size, and location of facial landmarks in the sketch and the ground truth can vary considerably. These situations reflect on the generator loss values and generally cause distortions that damage the realism of the rendering. To solve this problem, the sketches and photos are transformed so that the faces are aligned and the position of the eyes are consistent in each sketch-photo pair. This enabled higher photo-realism in the generated renderings. Additionally, the accuracy of the matching process is also improved [16].

3.3 Training

The model was trained on two experimental settings. The first one followed the aforementioned loss function to promote both an intermediate realistic face

[3] DeOldify API. Available on: https://github.com/jantic/DeOldify.

rendering and high matching accuracy. Here, the model was trained during 780, 1500 and 2 epochs, for CUFSF, CUFS and CelebA respectively, with batch size 8, using the Adam optimizer with a learning rate of 2×10^{-4} and parameter $beta_1 = 0.5$. The loss parameters λ_1 and λ_2 were experimentally set to 100 and 1, respectively. This allowed for a good balance of all the losses and increased the quality of the generated images. The pix2pix paper [5] also recommended a value of 100 for the λ_1 parameter. The second setting studied the effect on performance if the realistic face rendering generation was disposable. To achieve this, the loss terms $\mathcal{L}_{cGAN}$ and $\mathcal{L}_{L1}$, that promotes the realistic intermediate representation, were removed. In this setting, the model was trained during 1320, 2800 and 2 epochs, for CUFSF, CUFS and CelebA respectively, with batch size 12, using the Adam optimizer with a learning rate of 1×10^{-3} and parameter $beta_1 = 0.9$. To improve the robustness of the model and to avoid overfitting, data augmentation was applied to each pair of images. These were randomly cropped and horizontally mirrored before each epoch.

4 Results and Discussion

To evaluate the matching performance of the proposed model, its rank-N accuracy was computed on the test sets of CUFS, CUFSF and CelebA. Rank-N accuracy measures the fraction of test instances where the true correspondence is successfully found among the N strongest predictions offered by the model. In this case, we consider $N \in \{1, 5, 10\}$. After training the model with the CUFSF dataset (1089 train pairs), 54% rank-1 accuracy (100 test pairs) was attained. Over all considered ranks, the matching accuracy of the proposed method is superior or aligned to the alternative methods, as presented in Table 1. However, when dismissing photo-realistic rendering generation, the matching accuracy increases significantly, showing a trade-off between intermediate representation realism and matching performance that could be tuned to fit the objectives of specific application scenarios.

Table 1. Matching accuracy on the CUFSF and CUFS datasets, using different methods to enhance the sketch (r.r.g.: realistic rendering generation).

Method	**CUFSF**			**CUFS**		
	Rank-1	*Rank-5*	*Rank-10*	*Rank-1*	*Rank-5*	*Rank-10*
Sketch	49	77	88	47	80	90
pix2pix	53	83	92	52	79	86
IPMFSPS [7]	**74**	94	97	**80**	**95**	**97**
HFFS2PS [1]	-	-	-	36	69	-
Proposed (with r.r.g)	54	87	96	44	77	86
Proposed (without r.r.g)	**74**	**98**	**99**	59	85	91

On the CUFS dataset, using 88 sketch-photo pairs for training and 100 pairs for testing, the proposed method attained 44% rank-1 matching accuracy (see Table 1). These performance results are below, but nevertheless aligned with the alternative methods evaluated on the same settings. When removing the rendering realism constraints in the loss, the accuracy at all ranks increases considerably, achieving much better performance.

Considering the significantly smaller size of the CUFS training set *vs.* the size of the training set of the CUFSF dataset, these results may denote that the proposed method is more sensitive to scarce data than the alternatives, failing to offer more general solutions. Furthermore, knowing the images in the CUFS dataset are much less diverse than those of CUFSF (regarding subject ethnicity, skin color, hair color, background), one can argue that the proposed method offers the greatest advantages over the alternatives in more challenging scenarios.

On the CelebA dataset, using 200599 sketch-photo pairs for training and 2000 test pairs, a rank-1 accuracy of 67% was attained (see Table 2). In order to calculate the accuracy we select one sketch-photo pair per identity and add it to the query set. We add the rest of the sketch-photo pairs to the gallery. Then we generate a realistic representation of every sketch in the query set, and order the identities in the gallery by their average distance to the generated image. We use the ordered list of identities to calculate the rank-N accuracy. Since the process of selecting the query sketch-photo pairs is random we repeat this process 10 times and take the average.

4.1 Realistic Generation Performance

After the evaluation of matching performance, the realism of the intermediate representations generated by the proposed method were evaluated. Examples of these photo-realistic images and the corresponding sketches and ground-truth photos from the CUFSF, CUFS and CelebA datasets can be seen in Fig. 3.

Measuring realism is a difficult task, since the concept is highly subjective. However, we can assume that, if a generated image is sufficiently similar to the corresponding ground-truth, it is realistic. Hence, we use similarity/dissimilarity

Table 2. Matching accuracy on the CelebA dataset, using different methods to enhance the sketch (r.r.g.: realistic rendering generation).

Method	**Accuracy (%)**		
	Rank-1	*Rank-5*	*Rank-10*
Sketch	18	33	43
QGS2PIS [10]	66	80	**92**
FAGDS2PS [6]	66	-	-
Proposed (with r.r.g.)	63	81	86
Proposed (without r.r.g)	**67**	**83**	88

metrics that are common in related literature works for an objective evaluation, along with a visual subjective inspection of the test results. These metrics were the Fréchet Inception Distance (FID) [3], the Inception Score (IS) [15], and the Structural Similarity Index (SSIM) [20].

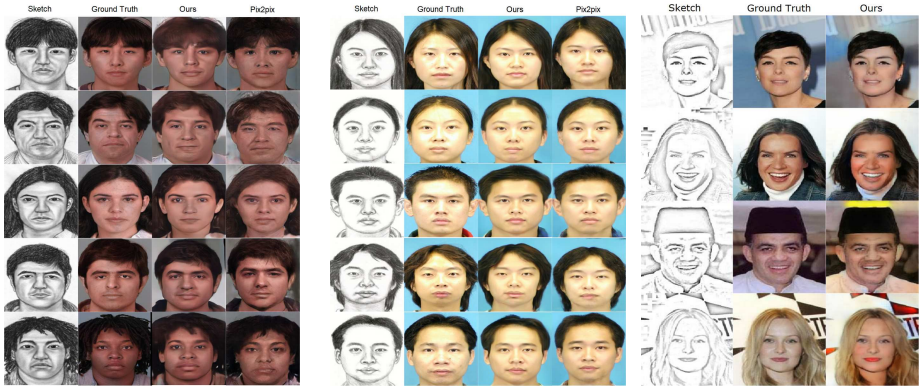

Fig. 3. Images generated by our method using the CUFSF dataset (on the left); Images generated by our method using the CUFS dataset (on the middle); Images generated by our method using the CelebA dataset (sketches were generated using an edge detection algorithm) (on the right).

As expected, the results through the similarity/dissimilarity metrics are better when the proposed method includes realistic rendering generation (see Table 3), as the method has learnt to generate images that are similar to the respective ground-truth photos. With CUFSF data, the results with the proposed method are slightly inferior but comparable to the alternatives, once again showing that the proposed method finds advantages in more challenging data.

Visually, we can confirm that the results of the proposed method are, indeed, very similar to the alternative method (see Fig. 3). Considering the improvement in performance, maintaining the degree of photo-realism is a positive aspect of

Table 3. Comparison between different methods to enhance the sketch, using the CUFSF and CUFS datasets (r.r.g.: realistic rendering generation).

Method	**CUFSF**			**CUFS**		
	FID ↓	IS ↑	SSIM ↑	FID ↓	IS ↑	SSIM ↑
Ground Truth	0.0	1.71	1.0	0.0	1.39	1.0
pix2pix	**70.54**	**1.69**	0.59	**41.46**	1.26	0.62
HFFS2PS [1]	-	-	-	58.50	1.43	**0.70**
Proposed (with r.r.g)	83.51	1.47	**0.60**	74.74	**1.54**	0.61
Proposed (without r.r.g)	330.41	1.48	0.32	321.96	1.35	0.45

Fig. 4. Intermediate representation of the sketches using the CUFSF dataset (on the left); Intermediate representation of the sketches using the CUFS dataset (on the right).

the proposed method. Furthermore, the proposed method delivers, in most cases, images that retain most information needed to identify the person represented in the sketch. Nevertheless, the method presents a worrying lack of diversity in specific facial features (such as hair, skin, or eye color) that should be addressed. In fact, when trained without photo-realistic generation, the method avoids these details by using unrealistic color schemes (see Fig. 4). To improve this shortcoming, an approach similar to the one proposed in [4], which consists in giving facial attribute information along with the sketch, could be an important addition to the proposed model.

5 Conclusion

This paper proposes an end-to-end deep method for sketch-to-photo matching that promotes the generation of a photo-realistic intermediate representation of the face depicted on the input sketch. As an end-to-end model, jointly trainable, it aims to eliminate performance limitations associated with separately optimized processes. Upon evaluation, the matching process showed performance improvements over the state-of-the-art and the generation of face renderings offered realistic results. When disregarding realistic rendering generation, the performance results improved. Despite the promising results, further efforts should be devoted to improve the face rendering generation component. Namely, the limited diversity of face characteristics like hair, eyes, and skin color on generated images should be addressed, in order to improve both the realism of the results and the matching performance.

Acknowledgements

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project UIDB/50014/2020, and within the PhD grant “SFRH/BD/137720/2018”. Portions of the research in this paper use the FERET database of facial images collected under the FERET program, sponsored by the DOD Counterdrug Technology Development Program Office.

References

1. Chao, W., Chang, L., Wang, X., Cheng, J., Deng, X., Duan, F.: High-fidelity face sketch-to-photo synthesis using generative adversarial network. In: ICIP. pp. 4699–4703 (2019)
2. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
3. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS, pp. 6626–6637 (2017)
4. Iranmanesh, S.M., Kazemi, H., Soleymani, S., Dabouei, A., Nasrabadi, N.M.: Deep sketch-photo face recognition assisted by facial attributes. In: IEEE BTAS. pp. 1–10 (2018)
5. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017)
6. Kazemi, H., Iranmanesh, M., Dabouei, A., Soleymani, S., M. Nasrabadi, N.: Facial attributes guided deep sketch-to-photo synthesis. In: WACVW. pp. 1–8 (2018)
7. Lin, Y., Ling, S., Fu, K., Cheng, P.: An identity-preserved model for face sketch-photo synthesis. IEEE Signal Processing Letters **27**, 1095–1099 (2020)
8. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV (2015)
9. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv (2014)
10. Osahor, U., Kazemi, H., Dabouei, A., Nasrabadi, N.: Quality guided sketch-to-photo image synthesis. arXiv (2020), 2005.02133
11. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: British Machine Vision Conference (2015)
12. Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J.: The FERET Evaluation Methodology for Face Recognition Algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**, 1090–1104 (2000)
13. Phillips, P.J., Wechsler, H., Huang, J., Rauss, P.: The FERET database and evaluation procedure for face recognition algorithms. Image and Vision Computing Journal **16**(5), 295–306 (1998)
14. Pramanik, S., Bhattacharjee, D.D.: An approach: Modality reduction and face-sketch recognition. arXiv (2013)
15. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., Chen, X.: Improved techniques for training gans. In: NeurIPS, pp. 2234–2242 (2016)
16. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. CVPR (Jun 2015)
17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)
18. Wang, M., Deng, W.: Deep face recognition: A survey. arXiv (2018)
19. Wang, X., Tang, X.: Face photo-sketch synthesis and recognition. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 31. IEEE (2009)
20. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
21. Zhang, W., Wang, X., Tang, X.: Coupled information-theoretic encoding for face photo-sketch recognition. In: CVPR (2011)