Title for citation

SCIENCE CHINA
Information Sciences

• Supplementary File •

Face-Sketch Learning with Human Sketch-Drawing Order Enforcement

Liang Chang[1], Lihua Jin[1], Lifen Weng[2], Wentao Chao[3], Xuguang Wang[3], Xiaoming Deng[4*] & Qiulei Dong[5,6,7*]

[1]*School of Artificial Intelligence, Beijing Normal University, Beijing* 100875*, China;*
[2]*Department of Design Art, Xiamen University of Technology, Xiamen, Fujian* 361024*, China;*
[3]*Department of Automation, North China Electric Power University, Baoding* 071003*, China;*
[4]*Beijing Key Laboratory of Human Computer Interactions, Institute of Software, Chinese Academy of Sciences, Beijing* 100190*, China;*
[5]*National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing* 100190*, China;*
[6]*University of Chinese Academy of Sciences, Beijing* 100049*, China;*
[7]*Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Beijing* 100190*, China*

Appendix A Details of the Order Enforced Photo-to-Sketch Dataset

The Ord-Sketch dataset used in the present study contained 400 face photos and 10000 corresponding sketches. To obtain the sketches to be uploaded in the dataset, three painters were asked to draw them based on the provided face photos following the unified drawing procedure. During this procedure, each painter created a sequence of 25 sketches with incremental degrees of fineness for each photo sequentially utilizing a graphic pen tablet Wacom DTK-1301/K0-F following the same collecting order.

Appendix B Network Structure and Training Details

At each stage, discriminators were utilized to estimate whether an input sketch was true or fake. It consisted of five convolution layers. The discriminators at the first three layers used the filters with the kernel size of 4×4 and the stride of 2×2, while those at the last two layers employed the filters with the stride of 1×1. The basic structural combination of a discriminator was $conv - bn - res - lrelu$, with the activation function utilized in the last layer was *sigmoid*. The utilized training strategy was described in detail in Algorithm B1.

Appendix C Experiments

In the conducted experiments, we utilized the developed Ord-Sketch dataset to evaluate the performance of SO-Net and several representative methods considered as benchmarks, including LLE [3], MRF [4], RSLCR [18], and Pix2Pix [9].

Appendix C.1 Experimental Settings

To conduct the experiment, we randomly selected 88 face photos and their corresponding intermediate sketches for training in the developed order enforced photo-sketch database based on the CUHK database, and the remainder ones were utilized for testing. Concerning the benchmark methods, we employed the same photos and their corresponding complete sketches for training. All training and testing images were cropped into ones with the size of 256×256 pixels. The training images were processed by applying random cropping and flipping.

Appendix C.2 Face Sketch Synthesis

To perform a quantitative comparison in terms of performance estimates for all considered methods, we utilized four evaluation criteria: peak signal-to-noise ratio(PSNR), structural similarity index metrics(SSIM) [19], multiscale structural similarity(MS-SSIM) [20], and Manual ranking(MR). Considering that it was relatively subjective to evaluate whether a synthesized sketch was appropriate or not, a single metric (such as PSNR, SSIM, or MS-SSIM) was not capable of reflecting

* Corresponding author (email: xiaoming@iscas.ac.cn, qldong@nlpr.ia.ac.cn)

Algorithm B1 The training strategy for SO-Net

Require: Initializing SO-Net parameters;

1: **for** the number of training iterations **do**

2: Calculate the generator loss for each stage $i, i \in \{1, \cdots, 5\}$:

3: **if** $i = 1$, **then**

$$\mathcal{L}_G^i \Leftarrow \log(1 - D_i(G_i(Z|P)) + \lambda_1 \mathcal{L}_{rec}^i + \lambda_2 \mathcal{L}_{edge}^i$$

4: **else**

$$\mathcal{L}_G^i \Leftarrow \log(1 - D_i(G_i(Z_i|P, Y_{i-1}^{'})) + \lambda_1 \mathcal{L}_{rec}^i + \lambda_2 \mathcal{L}_{edge}^i$$

5: **end if**

6: Update the discriminator parameters of the network by ascending its gradient as follows:

$$\nabla_{\theta_D}(\mathcal{L}_D)$$

7: Update the generator parameters of the network by descending its gradient as follows:

$$\nabla_{\theta_G}(\mathcal{L}_G)$$

8: **end for**

Table C1 Quantitative evaluation of the proposed approach and several representative face-sketch synthesis methods.

Methods	LLE	MRF	RSLCR	Pix2Pix	SO-Net
PSNR	17.5713	16.1987	17.7020	16.6563	**18.6757**
SSIM	0.6074	0.5885	0.6326	0.6093	**0.6702**
MS-SSIM	0.7550	0.7034	**0.7692**	0.7488	0.7587
MR	3.0704	4.1833	2.6148	2.6444	**2.2944**

the overall quality of a synthesized sketch. Therefore, we asked 20 subjects to manually rank the synthesized sketches obtained by the five considered methods (namely, MR criterion was considered). During manual ranking, we provided the subjects with the original photos, ground-truth sketches, and the resulting sketches generated by the considered methods. The ranking score ranged from 1 to 5, where '1' corresponded to the best score and '5' to the worst one (assigning equal ranking scores was allowed).

Examples of the synthesized sketches obtained by the proposed method corresponding to the five processing stages are represented in Fig. C1 and Fig. C2. It can be seen that the obtained synthesized sketches demonstrate clear human face contours already at Stage 1. At Stage 2, recognizable facial features can be distinguished in the synthesized sketches. At Stage 3, the synthetic sketch represents more specific facial features, such as completely visible eye, nose, and mouth. At Stage 5, it is possible to see shadows in the generated sketches. Therefore, the synthesized sketches at different stages overall follow the order maintained by an artist during the process of sketch drawing.

Fig. C3 and Fig. C4 represent the synthesized example sketches generated by the considered methods. It can be observed that the sketches obtained by LLE, MRF, and RSLCR have blur artifacts and lack shading in facial region, while that corresponding to Pix2Pix contains minor shading but lacks local sharpness contrast. Analyzing the resulting sketches generated by the proposed method, we conclude that it can successfully derive high-quality face sketches with clear face contours and smooth texture.

In addition, Table C1 outlies the scores in terms of PSNR, SSIM, MS-SSIM, and MR corresponding to the synthesized sketches for all considered methods. It can seen that the proposed SO-Net approach achieves the best SSIM and MS-SSIM scores, as well as the comparable PSNR score although being slightly lower than those of LLE and RSLCR. Furthermore, the proposed SO-Net obtains the best MR score, 2.2944 among all alternative methods, indicating that the sketches generated by SO-Net conform with the people's aesthetic sense of taste to a greater extent.

Moreover, we tested the proposed approach and the proposed methods utilizing the face photos obtained from CK+ [23], CUFS [4], and Wiki [22] datasets, as represented in Fig. C5 and Fig. C6. It can be seen that the proposed approach overall performs better when applied to these new datasets compared with the alternative methods. Although it performed slightly worse in the case of the Wiki dataset, it could be explained by the fact that the photos from the Wiki dataset became rather small after face registration and resize.

Appendix C.3 Sketch Recognition

In this subsection, we present the results of the implementation of null space linear discriminant analysis(NLDA) [21] conducted to evaluate the performance estimates of all considered sketch synthesis methods utilized for sketch recognition.

Here, we utilized 100 photo-sketch pairs for sketch recognition, so that 40 of them were used for training and the remaining

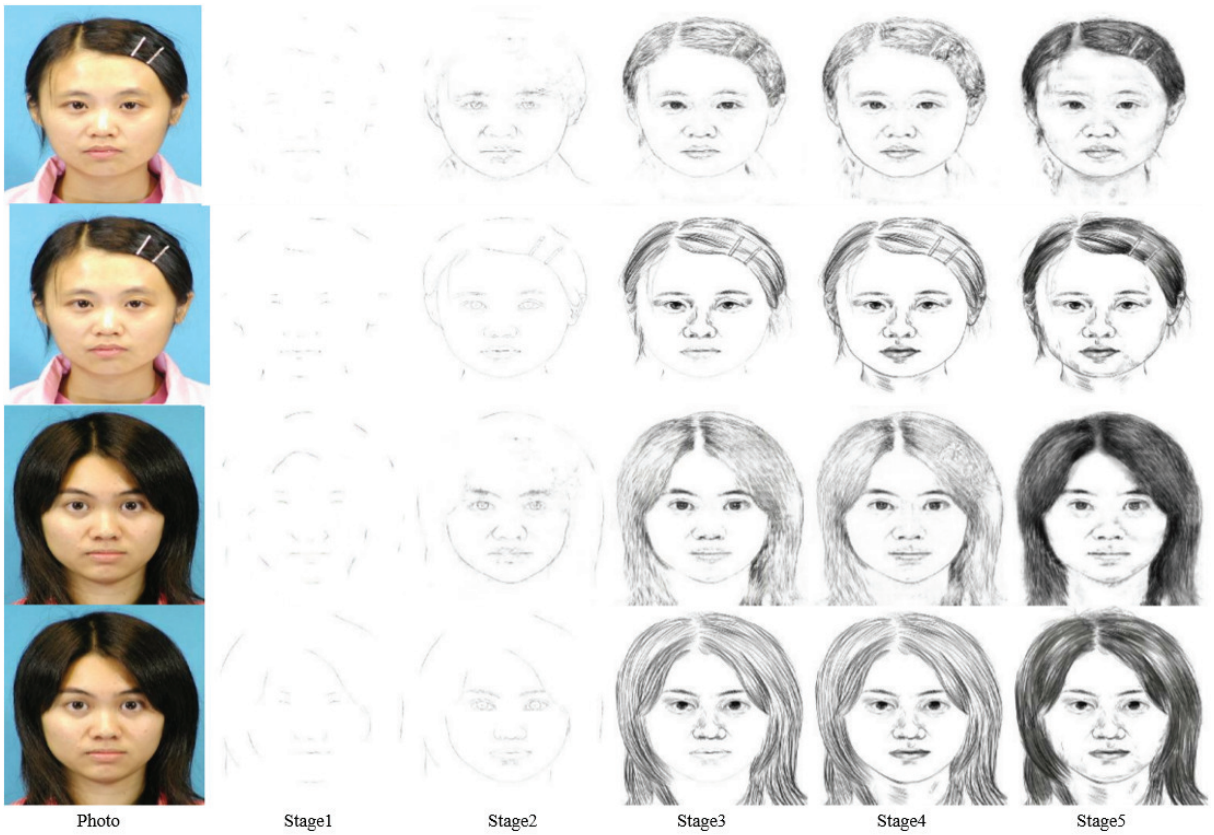

Figure C1 Examples of synthesized sketches from Chinese identities at the five stages in SO-Net. The first and third rows represent the synthesized sketches, and the second and fourth rows show the corresponding sketches drawn by painters.

60 ones for testing. Dimension reduction was performed by means of NLDA so that extracted feature vectors of a real photo and its corresponding synthesized sketch matched appropriately. The sketch recognition experiments were repeated 20 times with random sampling of the training and testing sets. The average recognition accuracies of all considered methods were calculated. The corresponding results are reported in Table C2. Here, it can be seen that the proposed SO-Net achieved better performance compared with the alternative methods.

Table C2 Comparison of the average recognition accuracy (in terms of NLDA) corresponding the synthesized sketch for all considered methods.

Methods	LLE [3]	MRF [4]	RSLCR [18]	Pix2Pix [9]	SO-Net
NLDA	0.8843	0.8617	0.8864	0.8914	**0.8942**

In addition, we tested all considered methods using different numbers $\{5, 10, ..., 35\}$ of elements in the training dataset. Fig. C7 represents the corresponding results for all considered methods. As shown in this figure, the proposed method outperformed the alternative ones in all cases corresponding to training datasets of different sizes. This indirectly confirmed the effectiveness of the proposed method for face synthesis.

Appendix C.4 Ablation Study

In addition, we conducted the experiments implying different combinations of stages. In Table C3, the results of quantitative comparison corresponding to different combinations of stages are presented. It can be seen that PSNR achieves the highest value for all stages. Moreover, the addition of the third stage resulted in a significant increase in the value of MS-SSIM and SSIM, which indicated that the third stage played an important role in the process of face sketch synthesis. Table C4 provides the results of comparing the average face recognition accuracy (in terms of NLDA) with those obtained for combinations of different stages. It can be seen that the combination of the first, second, and fifth stages resulted in achieving the highest accuracy (0.9401).

We analyzed the performance of the proposed method in terms of the accuracy of synthesized sketches at different stages of the face recognition process. Specifically, we compared the face recognition accuracy of the synthesized sketches obtained at different stages. From Table C5, it can be seen that at the later stages, higher face recognition rates were achieved. For

Figure C2 Examples of the synthesized sketches for non-Chinese identities corresponding the five stages in SO-Net. The first and third rows represent the synthesized sketches obtained by the proposed method, and the second and fourth rows show the corresponding sketches drawn by the painters.

Table C3 Evaluation metrics for combinations of different stages.

Combinations of Stages	5	1,5	1,2,5	1,3,5	1,2,3,5	1,3,4,5	1,2,3,4,5
PSNR	16.5097	16.7863	16.6751	16.9108	16.7584	16.7615	**16.9313**
MS-SSIM	0.7436	0.7530	0.7517	**0.7618**	0.7544	0.7508	0.7572
SSIM	0.6021	0.6150	0.6137	**0.6238**	0.6133	0.6105	0.6121

example, the recognition accuracies at the third, fourth, and fifth stages were 0.8627, 0.8968, and 0.9447, respectively. We concluded that it was consistent with the human visual perception, meaning that the facial features became more visible and identifiable as the number of stages increased.

Moreover, we performed a fair comparison between the proposed method and Pix2Pix. Here, we utilized all data, including the intermediate sketches, as input into the proposed method and Pix2Pix. The synthesis results were compared for these two approaches, as represented in Fig. C8. The first row corresponds to the synthesis results obtained by the proposed method; the second row represents those derived by Pix2Pix, and the third one shows ground truth sketches drawn by the artists. It can seen that the proposed method demonstrates better performances.

Appendix C.5 Failure Cases

Several failure results outputted by the proposed method are represented in Fig. C9. In these examples, the proposed method failed to perform concerning such elements, as the area of chin, ears, glasses, and hair. These failures could have occurred due to overexposure of the input photo in the areas of chin and neck, the reflection of glasses, etc.

Figure C3 Synthesized sketches corresponding to the face photos for Chinese identities obtained by the considered methods: (a) face photo; (b) sketches by painters; (c) LLE [3]; (d) MRF [4]; (e) RSLCR [18]; (f) Pix2Pix [9]; (g) SO-Net.

Table C4 Comparison of average recognition accuracy (via NLDA) with randomly combined stages.

Combinations of Stages	5	1,5	1,2,5	1,3,5	1,2,3,5	1,3,4,5	1,2,3,4,5
NLDA	0.9335	0.9355	**0.9401**	0.9298	0.9312	0.9334	0.9361

Figure C4 Synthesized sketches corresponding to the face photos for non-Chinese identities obtained by all considered methods: (a) original face photo; (b) sketches by painters; (c) LLE [3]; (d) MRF [4]; (e) RSLCR [18]; (f) Pix2Pix [9]; (g) SO-Net.

Figure C5 Examples of synthesized sketches from the CK+ [23], CUFS [4], and Wiki [22] datasets obtained using the trained model for non-Chinese faces. (a) and (b) are the synthesis results for CK+, (c) and (d) are those for CUFS; (e) and (f) are those for Wiki. The Wiki outputs are slightly worse compared with others as the input face photos are fairly small.

Table C5 Average face recognition accuracy (in terms of NLDA) for an increasing number of combined stages.

Combinations of Stages	1	1,2	1,2,3	1,2,3,4	1,2,3,4,5
NLDA	0.5652	0.6641	0.8627	0.8968	**0.9447**

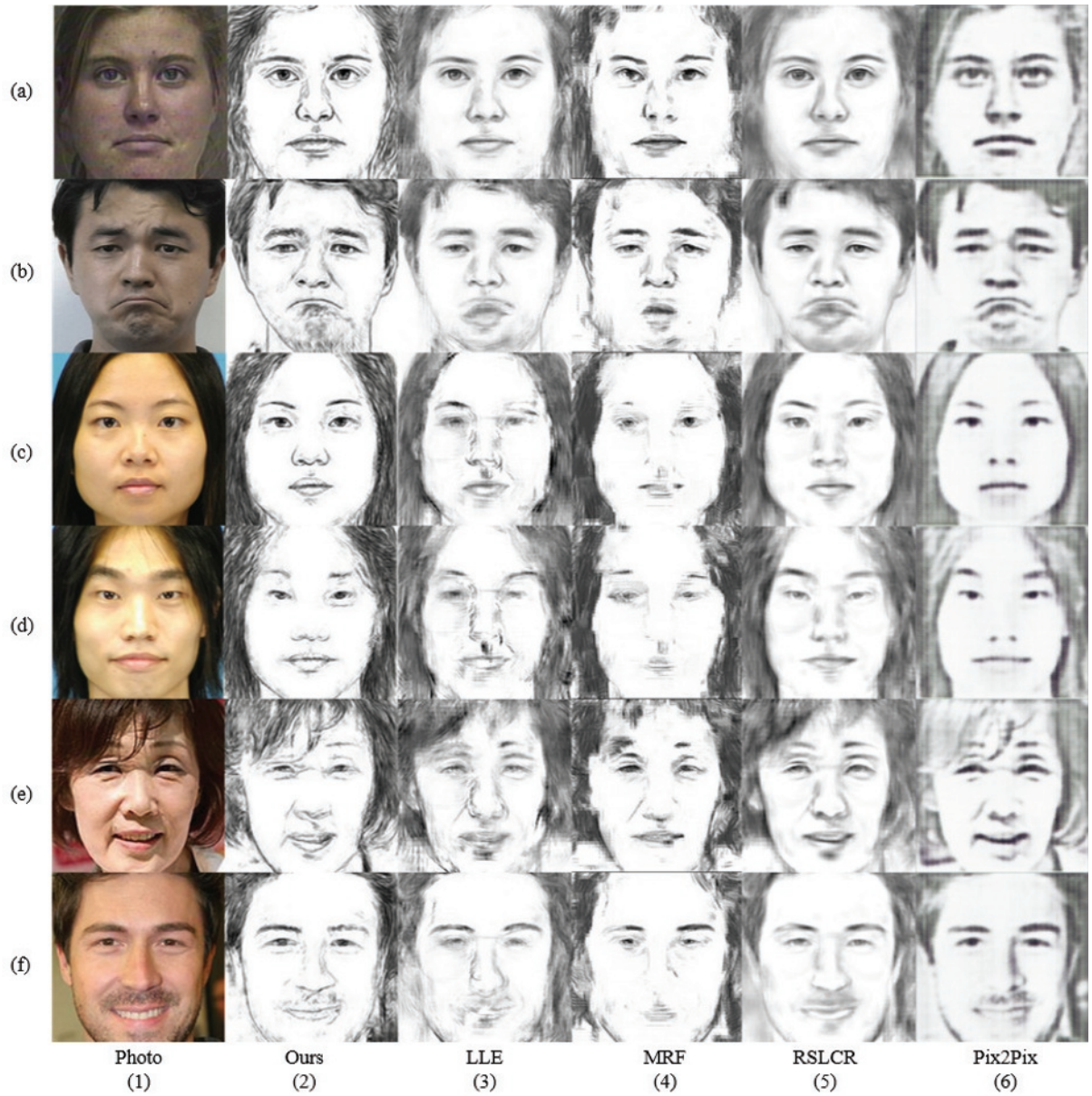

Figure C6 Examples of synthesized sketches corresponding to CK+ [23], CUFS [4], and Wiki [22] datasets using the considered methods. (a) and (b) are synthesis results for CK+ , (c) and (d) are those for CUFS, (e) and (f) are those for Wiki. Column (1) represents an original photo from the dataset, photos represented from column (2) to column (6) are the synthesis results obtained by the proposed method, LLE, MRF, RSLCR and Pix2Pix, respectively.

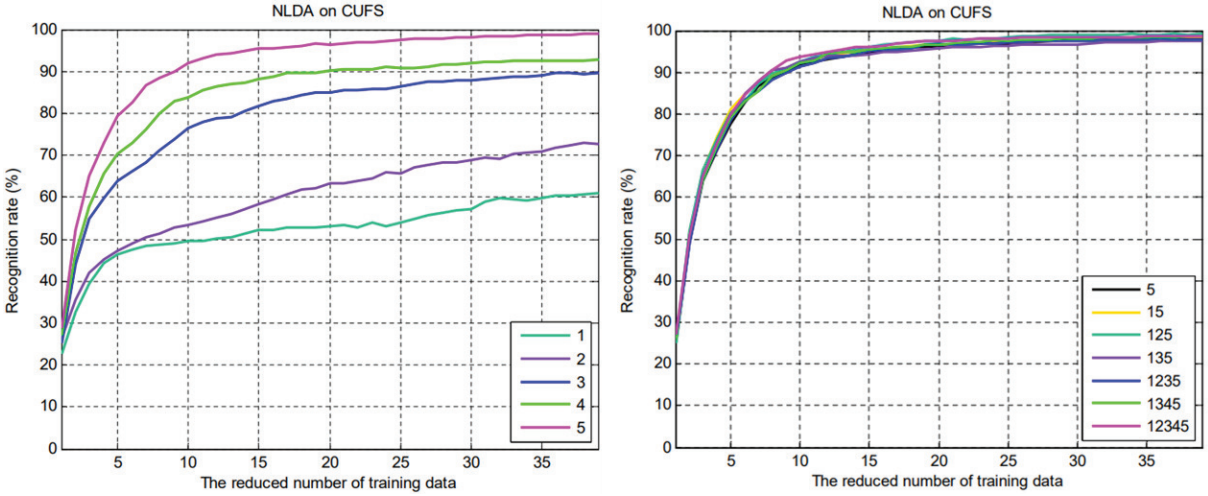

Figure C7 The graph on the left represents the results of face recognition accuracy comparison between all considered methods with a varying number of elements in the training dataset. The right graph depicts the results of face recognition accuracy comparison between different stages with a varying number of elements in the training dataset. It can be seen that the recognition accuracy augments according to an increase in the number of stages.

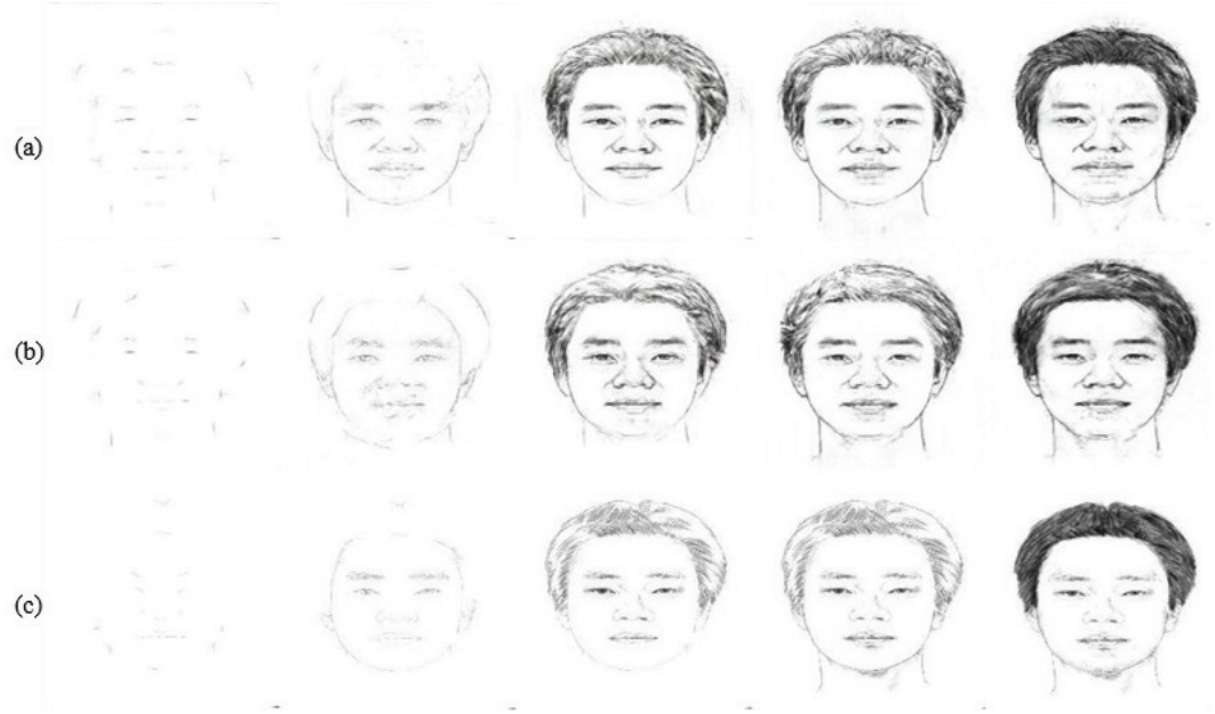

Figure C8 Comparison of the results corresponding to the five stages of the synthesis process. Figures from left to right represent the stages from Stage 1 to Stage 5. The first row provides the synthesis results obtained using the proposed method, the second row corresponds to those obtained by Pix2Pix, and the third row shows the ground truth sketches performed by artists.

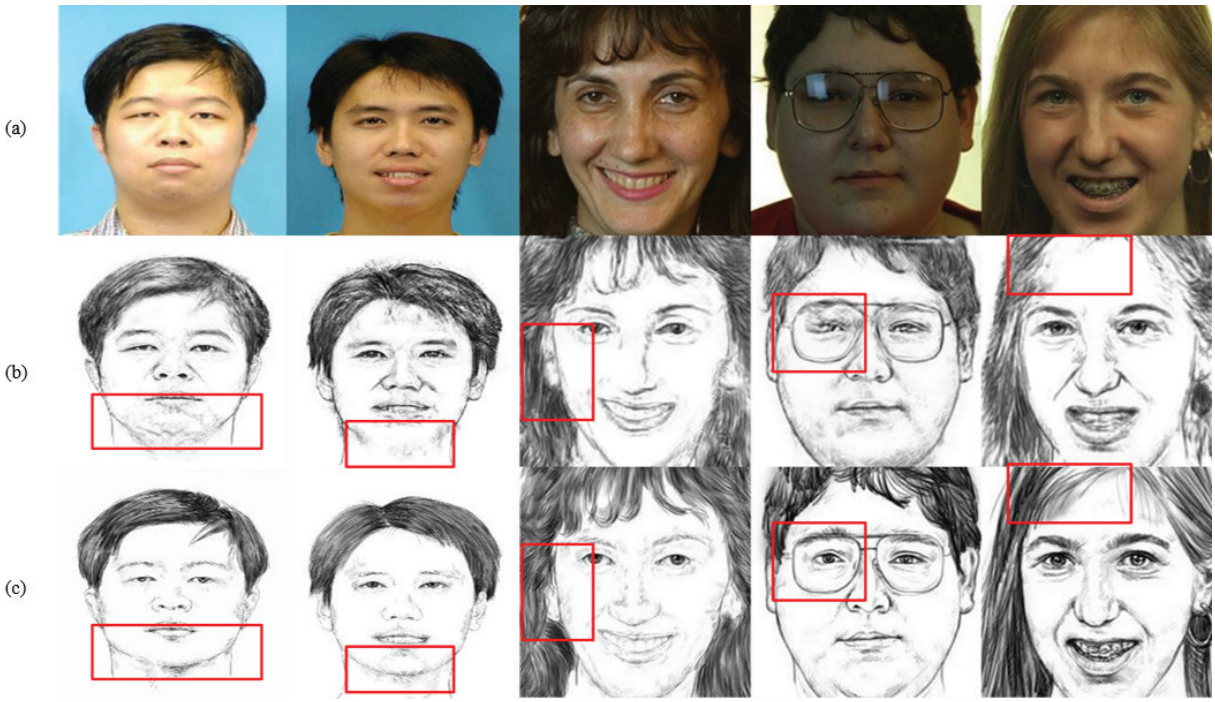

Figure C9 Several failure examples provided by the proposed method. Row (a) is the input photo; row (b) is the synthesized sketch from SO-Net; and row (c) is the ground truth sketch in the considered dataset. In these examples, the proposed method may fail concerning the area of chin, ears, glasses, or hair, which are highlighted with red rectangles.

References

1 Tang X, Wang X. Face Sketch Recognition. IEEE Transactions on Circuits and Systems for Video Technology, 2004, 14(1):50-57.
2 Tang X, Wang X. Face sketch synthesis and recognition, In: Proceedings of the 9th IEEE International Conference on Computer Vision (ICCV), Nice, France, 2003. 1:687-694
3 Liu Q, Tang X, Jin H, et al. A nonlinear approach for face sketch synthesis and recognition, In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 2005. 1:1005-1010
4 Wang X, Tang X. Face photo-sketch synthesis and recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(11):1955-1967
5 Zhou H, Kuang Z, Wong K K. Markov weight fields for face sketch synthesis, In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), RI, USA, 2012. 1091-1097
6 Chang L, Zhou M Q, Han Y, et al. Face sketch synthesis via sparse representation, In: Proceedings of the 20th International Conference on Pattern Recognition (ICPR), Istanbul,Turkey, 2010. 2146-2149
7 Wang N, Gao X, Tao D, et al, Face sketch-photo synthesis under multi-dictionary sparse representation framework, In: Proceedings of the 6th International Conference on Image and Graphics (ICIG), Hefei, Anhui, China, 2011. 82-87
8 Wang N, Li J, Tao D, et al. Heterogeneous image transformation. Pattern Recognition Letters, 2013, 34(1): 77-84
9 P. Isola, Zhu J, Zhou T, A. A. Efros. Image-to-image translation with conditional adversarial networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017. 5967-5976
10 M. Mirza, S. Osindero. Conditional generative adversarial nets. CoRR, 2014, abs/1411.1784
11 Yi Z, Zhang H, Tan P, Gong M. Dualgan: Unsupervised dual learning for image-to-image translation. In Proceedings of IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017. 2242-2251
12 Zhu J, Park T, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017. 2242-2251
13 Zhang L, Lin L, Wu X, Ding S, Zhang L. End-to-end photo-sketch generation via fully convolutional representation learning. In: Proceedings of the 5th ACM on International Conference on Multi-media Retrieval, Shanghai, China, 2015. 627-634
14 Wang L, V. Sindagi, V. M. Patel. High-quality facial photo-sketch synthesis using multi-adversarial networks. In: Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG), Xi'an, China, 2018. 83-90
15 Zhang S, Ji R, Hu J, Gao Y, et al. Robust face sketch synthesis via generative adversarial fusion of priors and parametric sigmoid. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI), Stockholm, Sweden, 2018. 1163-1169
16 Zhang W, Wang X, Tang X. Coupled information-theoretic encoding for face photo-sketch recognition. In: Proceedings of the 24th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 2011. 513-520
17 I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, et al. Generative adversarial nets. In:Annual Conference on Neural Information Processing Systems(NIPS), Montreal, Quebec, Canada, 2014. 2672-2680
18 Wang N, Gao X, Li J. Random sampling for fast face sketch synthesis. Pattern Recognition, 2018, 76:215-227.
19 Wang Z, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 2004, 13(4):600-612
20 Wang Z, E. P. Simoncelli, A. C. Bovik. Multiscale structural similarity for image quality assessment. In: Proceedings of the 37th Asilomar Conference on Signals, Systems & Computers, Pacific Grove, CA, USA, 2003. 2:1398-1402.
21 Chen L, Liao H, Ko M, Lin J, et al. A new lda-based face recognition system which can solve the small sample size problem. Pattern Recognition, 2000, 33(10):1713-1726
22 Rasmus Rothe, Radu Timofte, Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks.International Journal of Computer Vision (IJCV), July, 2016
23 P.Lucey, J. F. Cohn, T.Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The Extended Cohn-Kanade Dataset (CK+) A complete dataset for action unit and emotion-specified expression. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference, 2010. 94-101
24 Khaligh-Razavi S. M. and Kriegeskorte N. Deep supervised, but not unsupervised, models may explain IT cortex representation. PLOS Computational Biology, 2014, 10:1-29
25 Eickenberg M, Gramfort A, Varoquaux G, Thirion B. Seeing it all: Convolutional network layers map the function of the human visual system, NeuroImage, 2017, 152:184-194
26 Güçlü, U., van Gerven, MaJ. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. 2015,35:10005-10014
27 Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention(MICCAI), Springer, Cham, 2015. 9351
28 Zhang M, Wang N, Li Y, Gao X. Neural Probabilistic Graphical Model for Face Sketch Synthesis. IEEE Transactions on Neural Networks and Learning Systems, 2019, 1-15
29 Wang N, Gao X, Sun L, Li J. Anchored Neighborhood Index for Face Sketch Synthesis. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 2154-2163
30 Wang N, Gao X, Sun L, Li J. Bayesian Face Sketch Synthesis. IEEE Transactions on Image Processing, 2017, 26(3):1264-1274

31 Yang S, Qi Y, Qin H. Simultaneous structure and geometry detail completion based on interactive user sketches. Science China Information Sciences, 2012, 55(5): 1123-1137

32 Zhang Y, Chen X, Lin L, Xia C, Zou D. High-level representation sketch for video event retrieval. Science China Information Sciences, 2016, 59:1-15