Quality Metric Guided Portrait Line Drawing Generation from Unpaired Training Data (Appendix)

Ran Yi, Yong-Jin Liu, *Senior Member, IEEE,* Yu-Kun Lai, *Member, IEEE,* Paul L. Rosin

A1 Overview

This appendix includes the following material:

- detailed design of the network architecture (Section A2);
- more style examples in the training set (Section A3);
- three more ablation studies and their quantitative evaluation results (Section A4);
- all evaluation material used in the user study in Section 7.4 of the main paper (Section A5.1);
- comparison with APDrawingGAN++ (Section A5.2);
- more test results on other face dataset (Section A5.3).

A2 Details of Network Architecture

In the main paper, we summarize the flowchart of the network architecture in Figure 5 and introduce the architecture design principle in Section 4.2. Here we present the fine details of our proposed network architecture in Figure A7. We denote the output channel as c, convolution kernel size as k, and stride in a convolution layer as s. 'Norm' means the instance normalization layer and 'LReLU' means the leaky ReLU with $\alpha = 0.2$.

A3 More Style Examples in Training Set

In the main paper, we introduce the selected three representative styles from the collected data and show three examples in Figure 2: (1) the first style is from Yann Legendre and Charles Burns where thin parallel lines are used to draw shadows; (2) the second style is from Kathryn Rathke where few dark regions are used and facial features are drawn using simple flowing lines; (3) the third style is from vectorportal.com where continuous thick lines and large dark regions are utilized. Here we provide more examples in Figure A1.

- R. Yi is with BNRist, MOE-Key Laboratory of Pervasive Computing, the Department of Computer Science and Technology, Tsinghua University, Beijing, China; and the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China.
- Y.-J. Liu is with BNRist, MOE-Key Laboratory of Pervasive Computing, the Department of Computer Science and Technology, Tsinghua University, Beijing, China. Y.-J. Liu is the corresponding author. E-mail: liuyongjin@tsinghua.edu.cn.
- Y.-K. Lai and P.L. Rosin are with School of Computer Science and Informatics, Cardiff University, UK.

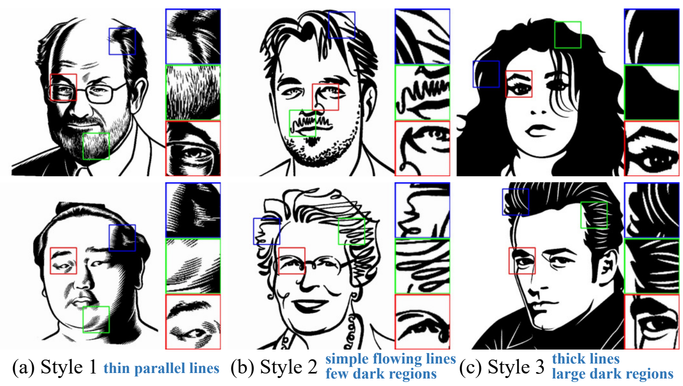

Fig. A1. More examples for three styles in the training set. Close-up views are shown alongside.

A4 Three More Ablation Studies

In Section 7.4 of the main paper, we study some key factors in our model, i.e., relaxed cycle-consistency loss, quality loss, local discriminators, and HED edge extraction. Here, we present three more ablation studies: (1) the first focuses on the style feature and style loss, (2) the second focuses on the truncation loss, and (3) the third focuses on how face region information is utilized in the discriminator.

In our method, when inputting a face photo and a style feature vector, the system outputs an APDrawing with style specified by the style feature vector. If we remove the style feature vector input and style loss from our system, when inputting a face photo, the model can output an APDrawing, but cannot generate APDrawings of different styles. Since the network is trained with mixed data, the output frequently exhibits different or mixed styles in different facial regions in an unpredictable way. Three examples are illustrated in Figure A2, in which all three photos contain a man face with beards. On the top of Figure A2(b), the generated APDrawing shows a parallel line style in the beard and hair regions (similar to style 1). In the middle of Figure A2(b), thick line and dark region style appears near the eyes, hair and jawline regions (similar to style 3). At the bottom of Figure A2(b), the generated APDrawing shows mixed styles. In comparison, as illustrated in Figures A2(c-e), after introducing style feature vector and style loss, our method can generate APDrawing results for each distinctive style, specified by the input style feature vector.

Fig. A2. Ablation study on style feature vector input and style loss. From left to right: input photos, results of removing style feature input and style loss, our results in styles 1, 2 and 3.

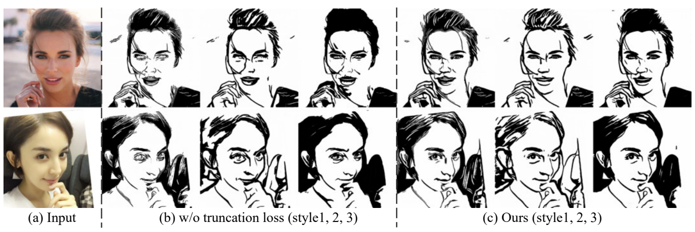

Fig. A3. Ablation study on truncation loss. From left to right: input photos, results of removing truncation loss (style1, 2, 3), and our results (style1, 2, 3).

We also study the role of truncation loss: two examples are shown in Figure A3. The truncation loss is designed to prevent the generated drawings from hiding information in small values. Without the truncation loss, the results sometimes do not draw full outlines of facial features (e.g., nose). As shown in Figure A3b, the nose in the first row lacks the middle outline and the nose in the second row lacks the right outline. In comparison, by adding the truncation loss, our system can generate complete outlines of different facial features.

We further perform a comparison by replacing local discriminators with a single discriminator which uses a new channel containing face region information. Our experiment shows that the results of this ablation are worse than those by our method, e.g., with partial facial features missing or messy (Figure A4). Also note that the face parsing masks are computed by an off-the-shelf face parsing network, with the parsed eyes/nose/lips regions dilated to make them cover the facial features. Some examples of the face parsing masks are shown in Figure A5. The results show that our system does not require accurate parsing masks.

The quantitative evaluation of the above ablation studies and comparisons are reported in Table A1. The FID values of these ablation studies are worse than ours:

- without style feature and style loss, the generated results are not of a uniform style, so the distance to each style is much larger than ours;
- without truncation loss, the FID also increases (worse).

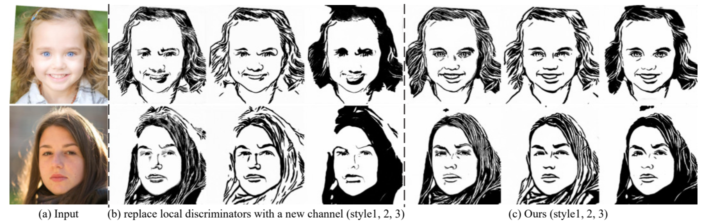

Fig. A4. Comparison of results with our local discriminators (c) and replacing them with a new channel (b) for input photos (a).

Fig. A5. Examples of face parsing masks.

These results show that the ablated components (style feature, truncation loss) are essential for our model. The comparison of replacing local discriminators with a single discriminator using a new channel has much larger FID value than ours, indicating a single discriminator using a new channel is harder to train, and our design of introducing local discriminators for important facial regions is more effective. Avg Δ shows the average difference between our method and each ablated version.

A5 More Results

A5.1 Material in the User Study

In Section 7.2 of the main paper, we compare our method with state-of-the-art methods in neural style transfer and image translation. In Section 7.3 of the main paper, we conduct a user study in which users sort the results of four methods (CycleGAN [1], ComboGAN [2], our conference version [3] and our method). Each time, users compare different methods' results of a single style. We denote 1 input photo and 4 generated drawings of a single style as a group. In total, 60 groups are evaluated in this user study. Among them, 20 groups are for style1 comparison, 20 groups are for style2 and 20 groups are for style3. We show all 60 groups in Figures A8-A13. For a more comprehensive comparison, we show results of all the 3 styles for the multi-modal methods (ComboGAN, our conference and ours) and highlight the compared group in the user study in green boxes. Note that all these 60 groups are randomly chosen from the test set. Our method outperforms the other three methods in most groups in terms of style similarity, face structure preservation and image visual quality. The results of the user study summarized in Section 7.3 of the main paper also demonstrate the advantage of our method, where 43.0% votes chose our method to be the best among the four methods, higher than the best vote percentages of the other three methods.

A5.2 Comparison with APDrawingGAN++

APDrawingGAN++ [4] is a deep neural network model specially designed for APDrawing generation by using a hierarchical structure and a distance transform loss. However, this method requires *paired* training data and cannot adapt well

TABLE A1
Fréchet Inception Distance (FID) of more ablation studies and comparisons. The FID values are computed between the set of generated APDrawings of each style and the collected true drawings of the corresponding style.

Methods	Style1 ↓	Style2 ↓	Style3 ↓	Avg Δ
w/o style feature	114.3	122.1	111.5	20.90
w/o truncation loss	81.7	120.1	99.2	5.27
replace D_{l*} with a single D	93.6	152.8	124.0	28.40
Ours	**81.2**	**114.3**	**89.7**	/

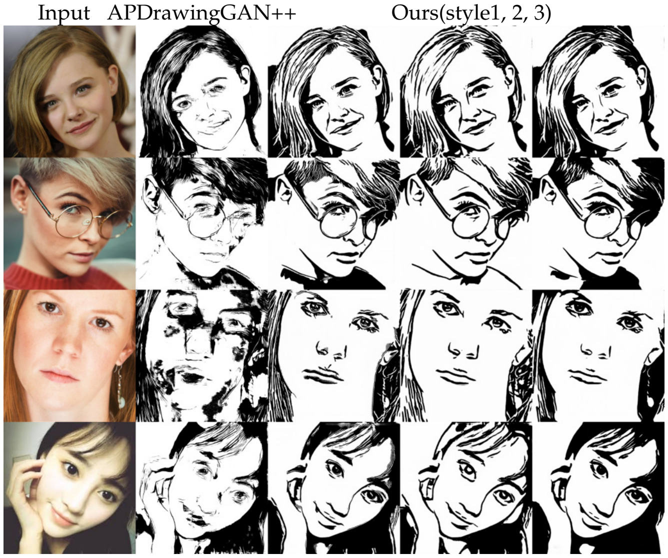

Fig. A6. Comparisons of APDrawingGAN++ and our method on challenging photos with arbitrary head orientation. From left to right: input photos, APDrawingGAN++ results, and our results (styles 1, 2, 3).

to face photos with unconstrained lighting in the wild due to the limited availability of paired training data. In comparison, our method only uses *unpaired* training data, which makes it possible to include more challenging photos into the training set. Therefore, our method can generate high quality APDrawings for challenging photos under various conditions. We compare the visual quality of APDrawingGAN++ and our method using some challenging examples as illustrated in Figure A14. These challenging examples include unconventional lighting conditions (1st-4th rows), unconventional expression or taking accessories like sunglasses (5th-7th rows), or blurry looking (8th-9th rows, zoom in to check). APDrawingGAN++ generates messy results for these challenging photos, while our method generates high-quality APDrawings with much better visual effect.

Moreover, APDrawingGAN++ uses a hierarchical network structure that feeds local rectangle regions around eyes, nose and mouth centers into local generators and discriminators. This setting cannot tolerate a large head tilt and requires that its input photos are in the upright orientation (i.e., the photo needs to be rotated so that the two eyes are on a horizontal line). Then the local regions of eyes, nose and mouth can be covered by rectangle regions. In comparison, although our model also has local discriminators, we use face masks (obtained from a face parsing network [5]), and the inputs to local discriminators are the masked eyes, nose, mouth regions. Therefore our method does not need the input images to be adjusted into the upright orientation. Comparisons of APDrawingGAN++ and our method on face photos with arbitrary head orientation are shown in Figure A6. The results show that APDrawingGAN++ often generates messy results and some boundaries of rectangle local regions are clearly visible, whereas our results are clean and have good visual quality.

A5.3 More Tests on the CelebAMask-HQ Dataset

In the main paper, we test our model on photos collected from Internet. Here, we further test our method on photos from the CelebAMask-HQ Dataset [6]. The results are summarized in Figure A15, showing that our method generates high quality results with good image and line quality on the CelebAMask-HQ Dataset.

References

[1] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251.

[2] A. Anoosheh, E. Agustsson, R. Timofte, and L. V. Gool, "ComboGAN: unrestrained scalability for image domain translation," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*, 2018, pp. 783–790.

[3] R. Yi, Y. Liu, Y. Lai, and P. L. Rosin, "Unpaired portrait drawing generation via asymmetric cycle mapping," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8214–8222.

[4] R. Yi, M. Xia, Y. Liu, Y. Lai, and P. L. Rosin, "Line drawings for face portraits from photos using global and local structure based GANs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, DOI (identifier) 10.1109/TPAMI.2020.2987931, 2020.

[5] S. Gu, J. Bao, H. Yang, D. Chen, F. Wen, and L. Yuan, "Mask-guided portrait editing with conditional GANs," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3436–3445.

[6] C. Lee, Z. Liu, L. Wu, and P. Luo, "MaskGAN: Towards diverse and interactive facial image manipulation," *CoRR*, vol. abs/1907.11922, 2019.

[7] P. L. Rosin, Y.-K. Lai, D. Mould, R. Yi, I. Berger, L. Doyle, S. Lee, C. Li, Y.-J. Liu, A. Semmo *et al.*, "NPRportrait 1.0: A three-level benchmark for non-photorealistic rendering of portraits," *Computational Visual Media*, 2021.

[8] 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," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.

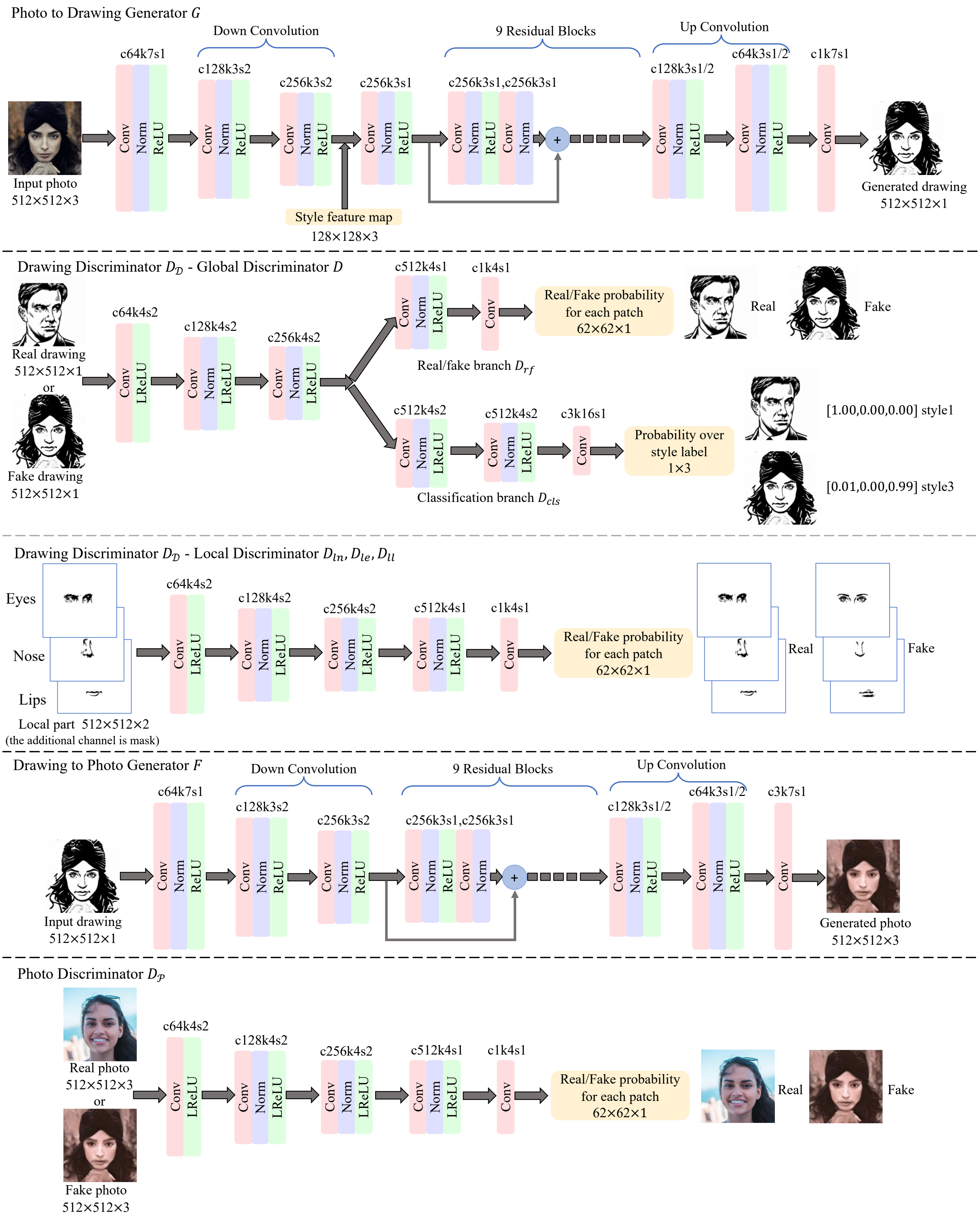

Fig. A7. Detailed network architecture of our model. We denote the output channel number as c, convolution kernel size as k, and stride in a convolution layer as s. 'Norm' means the instance normalization layer, and 'LReLU' means the leaky ReLU with $\alpha = 0.2$.

Fig. A8. More qualitative comparisons (user study material). From left to right: input face photos, CycleGAN [1] results, ComboGAN [2] results (style 1, 2, 3), results of our conference version (Ours-pre) [3] (style 1, 2, 3), and our results (style 1, 2, 3). In the user study, users compared each time the results of a single style. 60 groups are evaluated and there are 20 groups for each style. We show results of all the 3 styles and highlight the compared group in green boxes.

Fig. A9. More qualitative comparisons (user study material). From left to right: input face photos, CycleGAN [1] results, ComboGAN [2] results (style 1, 2, 3), results of our conference version (Ours-pre) [3] (style 1, 2, 3), and our results (style 1, 2, 3). In the user study, each time users compared results of a single style. 60 groups are evaluated and there are 20 groups for each style. We show results of all the 3 styles and highlight the compared group in green boxes.

Fig. A10. More qualitative comparisons (user study material). From left to right: input face photos, CycleGAN [1] results, ComboGAN [2] results (style 1, 2, 3), results of our conference version (Ours-pre) [3] (style 1, 2, 3), and our results (style 1, 2, 3). In the user study, users compared each time the results of a single style. 60 groups are evaluated and there are 20 groups for each style. We show results of all the 3 styles and highlight the compared group in green boxes.

Fig. A11. More qualitative comparisons (user study material). From left to right: input face photos, CycleGAN [1] results, ComboGAN [2] results (style 1, 2, 3), results of our conference version (Ours-pre) [3] (style 1, 2, 3), and our results (style 1, 2, 3). In the user study, users compared each time the results of a single style. 60 groups are evaluated and there are 20 groups for each style. We show results of all the 3 styles and highlight the compared group in green boxes.

Fig. A12. More qualitative comparisons (user study material). From left to right: input face photos, CycleGAN [1] results, ComboGAN [2] results (style 1, 2, 3), results of our conference version (Ours-pre) [3] (style 1, 2, 3), and our results (style 1, 2, 3). In the user study, each time users compared results of a single style. 60 groups are evaluated and there are 20 groups for each style. We show results of all the 3 styles and highlight the compared group in green boxes.

Fig. A13. More qualitative comparisons (user study material). From left to right: input face photos, CycleGAN [1] results, ComboGAN [2] results (style 1, 2, 3), results of our conference version (Ours-pre) [3] (style 1, 2, 3), and our results (style 1, 2, 3). In the user study, each time users compared results of a single style. 60 groups are evaluated and there are 20 groups for each style. We show results of all the 3 styles and highlight the compared group in green boxes.

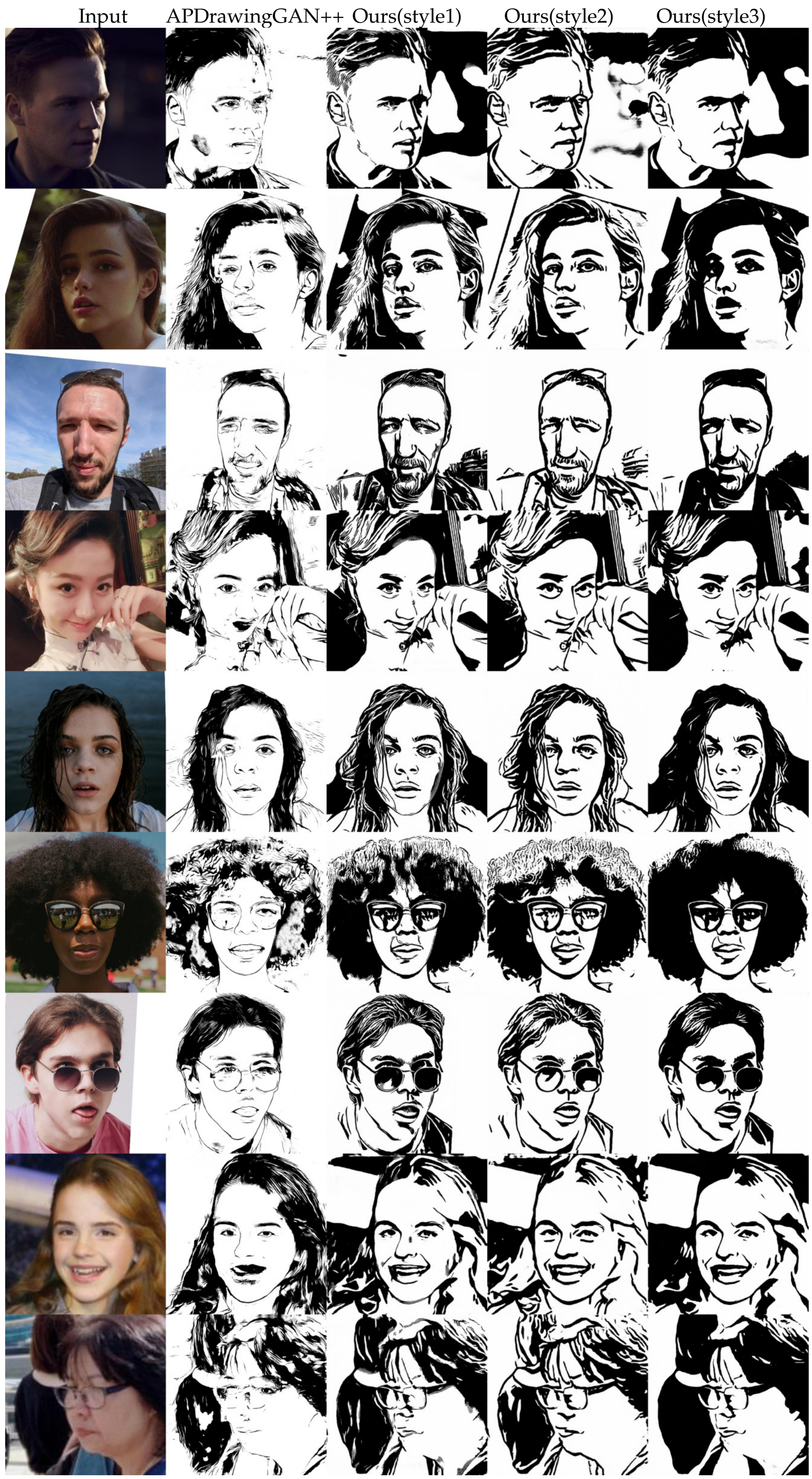

Fig. A14. Comparison of APDrawingGAN++ [4] and our method on face photos under some challenging situations. From left to right: input face photos, APDrawingGAN++ [4] results, our results (style1), our results (style2), our results (style3). The face photos in the 5-7th rows are from NPRportrait1.0 Benchmark [7]. The face photo in the 8th row is from LFW Dataset [8].

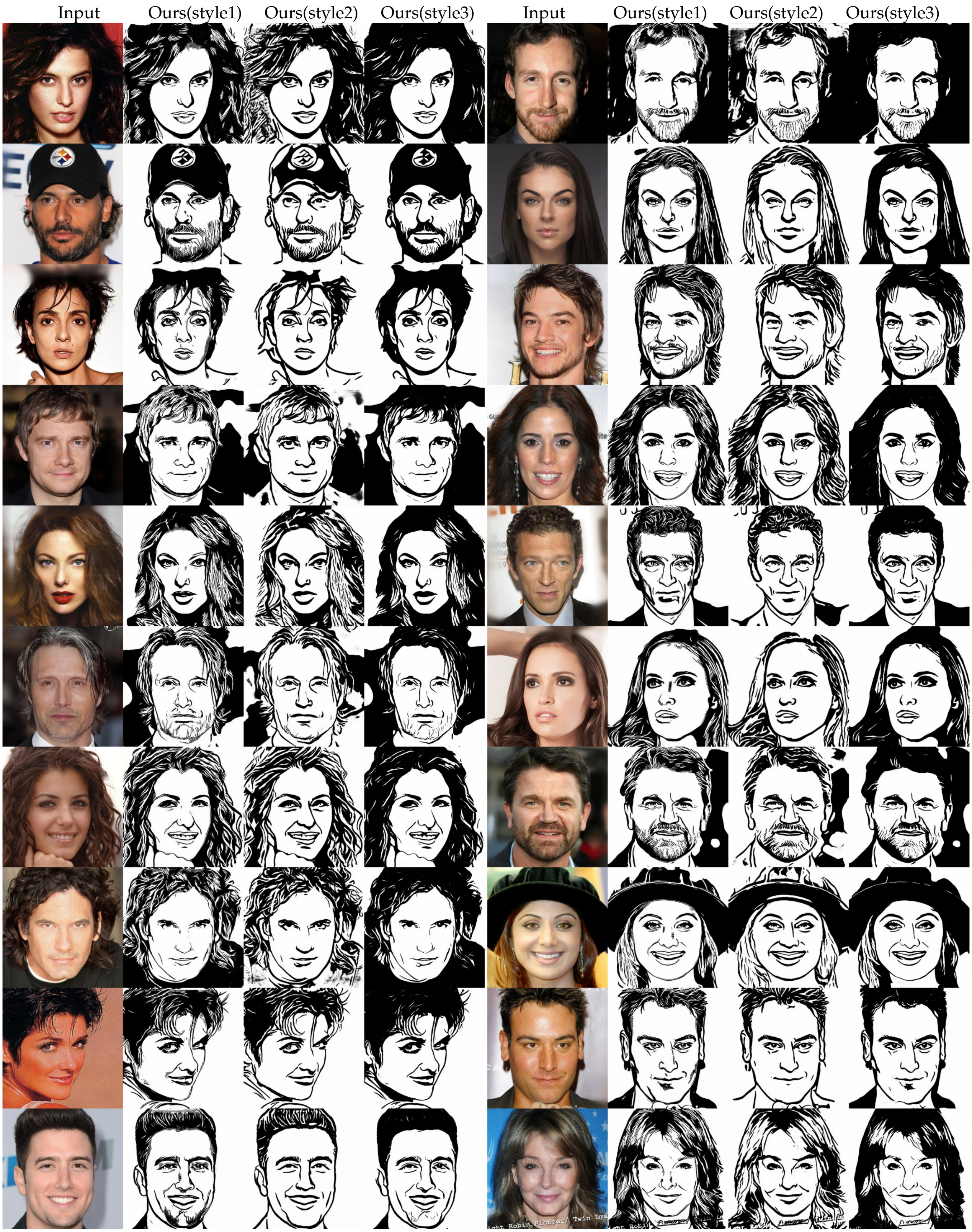

Fig. A15. More test results on CelebAMask-HQ Dataset [6]. From left to right: input face photos, our results (style1), our results (style2), our results (style3), input face photos, our results (style1), our results (style2), our results (style3).