Supplemental Materials DrawingInStyles: Portrait Image Generation and Editing with Spatially Conditioned StyleGAN

I. Data Preparation

We resort to the sketch extraction method in *DeepFaceDrawing* [1], which uses the PhotoCopy filter in PhotoShop and a sketch simplification method [2], to process the generated portrait images and thus get the sketch-image pairs. For the construction of the semantic map repository of the generated portrait images, we utilize an image segmentation network, BiSeNet [3], pre-trained on the CelebAMask-HQ dataset [4]. Finally, we get 45K (40K for training) data (image-sketch-semantic map-latent code) samples for training our *SC-StyleGAN* module.

To assist users of our system in sketching, we provide both the global face templates and the local component details in the suggestive sketching interface. To provide the global sketching guidance, we use a subset (17K samples) of the FFHQ dataset [5] and extract the corresponding sketches as *DeepFaceDrawing* [1] to construct the sketch template repository. We also predict the semantic maps of the 17K images using the same scheme mentioned above as the semantic map dataset and extract the contours of the semantic maps to provide a retrieval intermediate in the initial global sketching stage.

For the local detail suggestion, we choose to use the CelebAMask-HQ dataset [4] as the component sketch candidates, since CelebAMask-HQ contains a relatively larger volume (30K) of accurate semantic maps. We roughly divide each semantic mask into eight regions of interest, namely, "left eye", "right eye", "nose", "mouth", "facial skin", "glass", "hat", and "hairs" to embed the individual component region details (Figure 1). We resort to the same sketch extraction method as mentioned above to get the component sketches for constructing the component sketch repositories for the local refinement. To reduce the overlap between different semantic regions, we further extract each sketch image with guidance provided by a dilated region mask instead of cropping with a rectangle region [1], and record the corresponding sketch and semantic mask. In this way, we get clean component sketches, which reduce the conflict when replacing with the original component sketch. In addition, we record the portrait pose attribute ($[yaw, pitch, roll]$ triplets) of each image in both the FFHQ subset and CelebAMask-HQ dataset for pose selection.

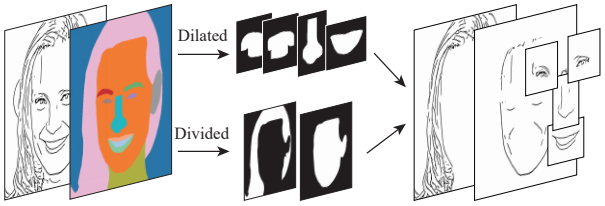

Fig. 1. An illustration of our local detail sketch extraction process. After getting a sketch from a real image, we decompose the sketch with respect to the dilated semantic masks of inner components ("left eye", "right eye", "nose" and "mouth" and "glass") to reduce the conflict among them when assembling. We extract the other component regions according to their original masks.

Fig. 2. Visual comparison on the side face generation between DeepFaceDrawing [1] (Middle) and our method (Right) given the same sketch input (Left). Here our method is trained with sketch only.

References

[1] S.-Y. Chen, W. Su, L. Gao, S. Xia, and H. Fu, "Deepfacedrawing: Deep generation of face images from sketches," *ACM Transactions on Graphics*, vol. 39, no. 4, pp. 72:1–72:16, 2020.

[2] E. Simo-Serra, S. Iizuka, K. Sasaki, and H. Ishikawa, "Learning to simplify: Fully convolutional networks for rough sketch cleanup," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 1–11, 2016.

[3] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 325–341.

[4] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," *arXiv preprint arXiv:1907.11922*, 2019.

[5] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4401–4410.

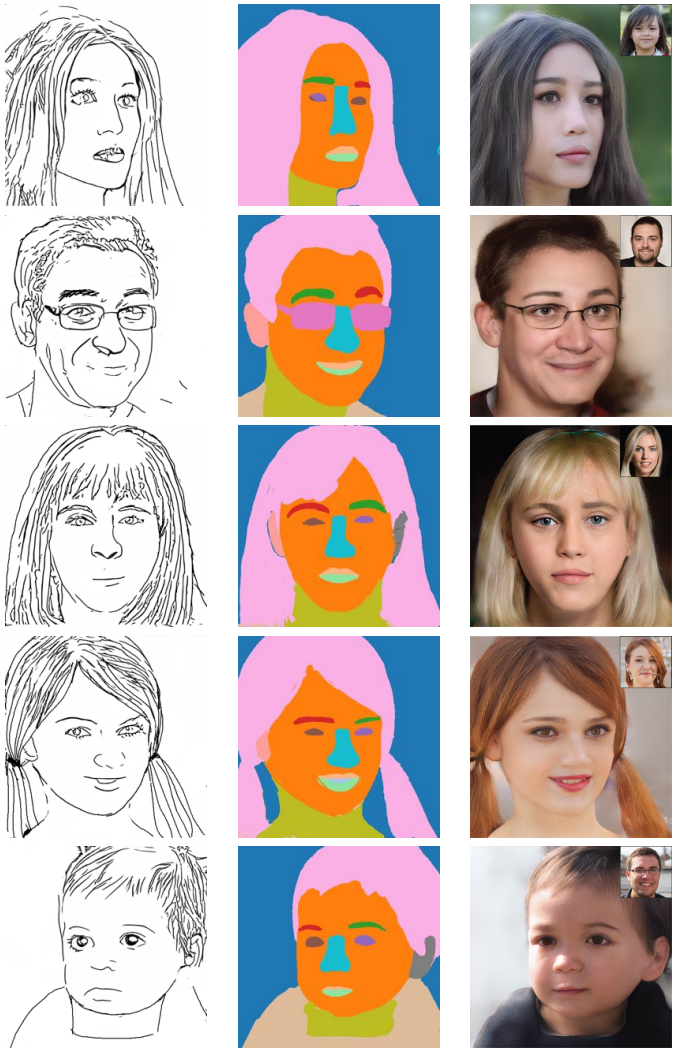

Fig. 3. Other results from the open-ended study. The left, middle, and right column of each group shows the final sketch, semantic map specified by the users, and the corresponding synthesized portrait image by our system. A reference style image is given in thumbnail on the upper right corner of each synthesized portrait image.