Supplemental Materials
DeepFaceDrawing: Deep Generation of Face Images from Sketches

SHU-YU CHEN*, Institute of Computing Technology, CAS and UCAS
WANCHAO SU*, School of Creative Media, City University of Hong Kong
LIN GAO†, Institute of Computing Technology, CAS and and UCAS
SHIHONG XIA, Institute of Computing Technology, CAS and and UCAS
HONGBO FU, School of Creative Media, City University of Hong Kong

1 OVERVIEW

In this supplemental material, we have listed the conducted experiments, additional gallery of the user study results, and more details of the implementation.

2 ABLATION STUDY OF K SELECTION

Fig. 1. The effect of selection of K in manifold projection. We set $K = 10$ in our implementation.

3 GALLERY OF THE USER STUDY RESULTS

We have listed more results obtained from the usability study in Figures 2 and 3.

4 SKETCHES USED IN THE PERCEPTIVE EVALUATION STUDY

In this section, we demonstrate the input sketches used in the perceptive evaluation in Figure 4.

5 COMPARISON OF DIFFERENT METHODS WITH TEST SET INPUT

We illustrate the visual comparison given the test set sketches (i.e., the edge maps) as input in Figure 5.

6 ABLATION STUDY OF THE SELECTION OF THE FEATURE DIMENSION

We show the MSE metrics of the reconstructed sketch with different latent feature dimensions. The quantitative results are elaborated in Table 1.

*Authors contributed equally.
†Corresponding author.

Authors' addresses: Shu-Yu Chen, Lin Gao and Shihong Xia are with the Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences. Wanchao Su and Hongbo Fu are with the School of Creative Media, City University of Hong Kong. Authors' e-mails: chenshuyu@ict.ac.cn, wanchao.su@my.cityu.edu.hk, gaolin@ict.ac.cn, xsh@ict.ac.cn, hongbofu@cityu.edu.hk.

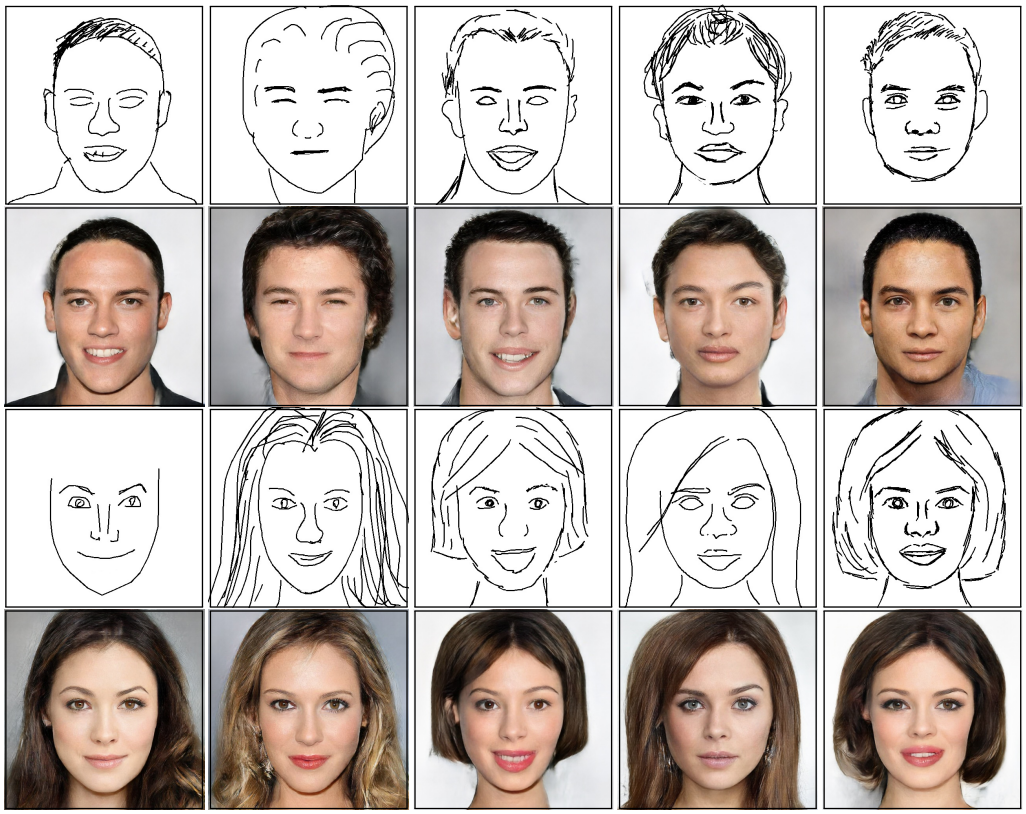

Fig. 2. More results: Part I.

Latent Dim	left-eye	right-eye	nose	mouth
128	0.370	0.323	0.423	0.359
256	0.247	0.228	0.288	0.287
512	**0.131**	**0.126**	**0.177**	**0.197**

Table 1. An ablation study for the number of feature dimensions in Component Embendding. Comparing the value of the MSE loss for each component in different feature dimensions, the loss decreases with the increase of the number of feature dimensions.

7 NETWORK ARCHITECTURES

Our network consists of Component Embedding, Feature Mapping Module and Image Synthesis Module. In training, we use Adam optimizer [Kingma and Ba 2014] with $\beta 1 = 0.5$ and $\beta 2 = 0.999$ and the initial learning rate is 0.0002.

7.1 Component Embedding Architectures

Our Component Embedding module contains five auto-encoders. Each auto-encoder consists of five encoding layers (Table 2) and five decoding layers (Table 3). The encoding layer is shown in Table 2. We add a fully connected layer in the middle to ensure the latent descriptor is of 512

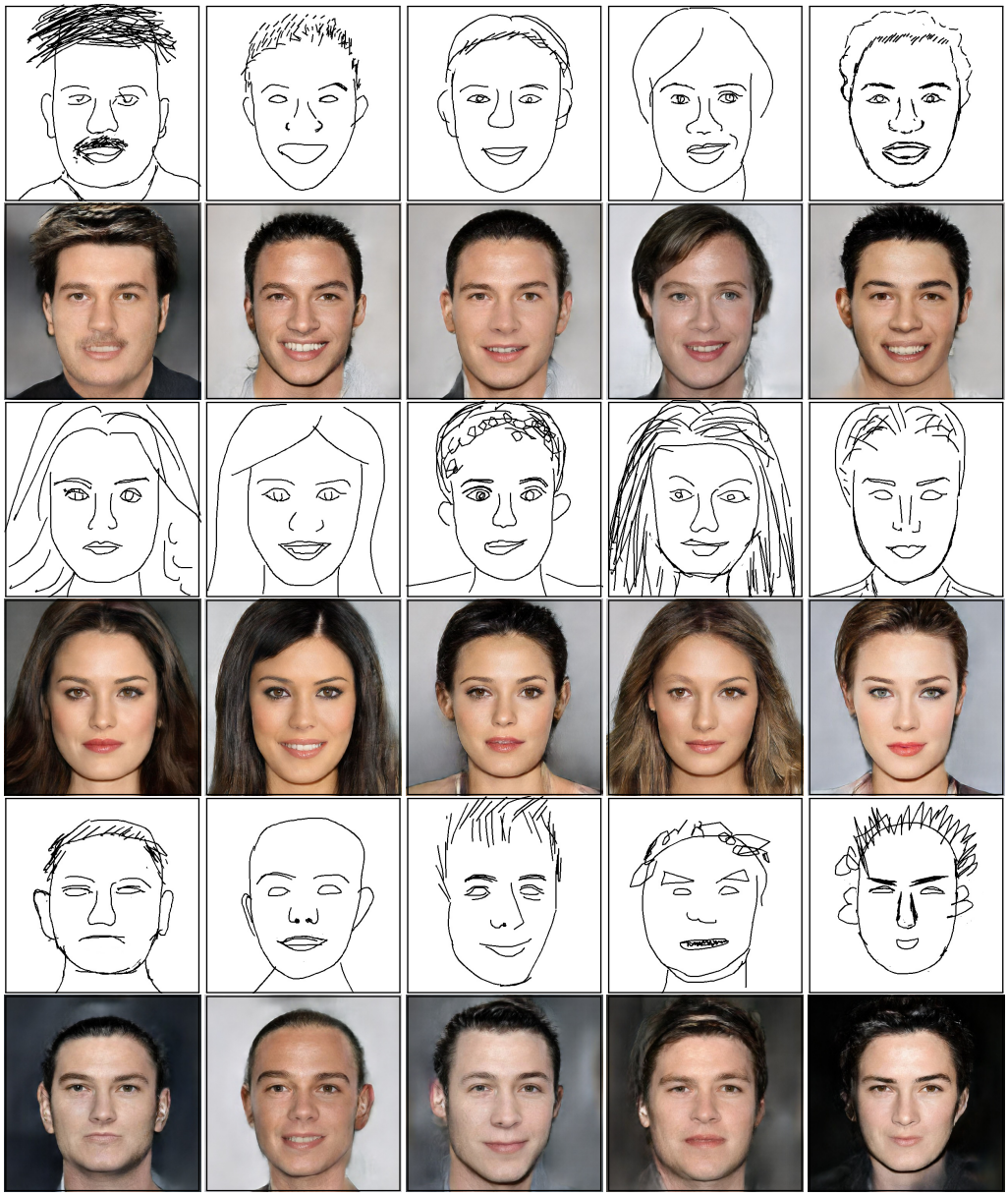

Fig. 3. More results: Part II.

dimensions for all the five components. The details of Component Embedding module are shown in Table 4.

7.2 Feature Matching Architectures

The Feature Matching module contains five decoding networks, which take as input the compact feature vectors obtained from the component manifolds and convert them to the corresponding size of feature maps for subsequent generation. See the details in Table 5.

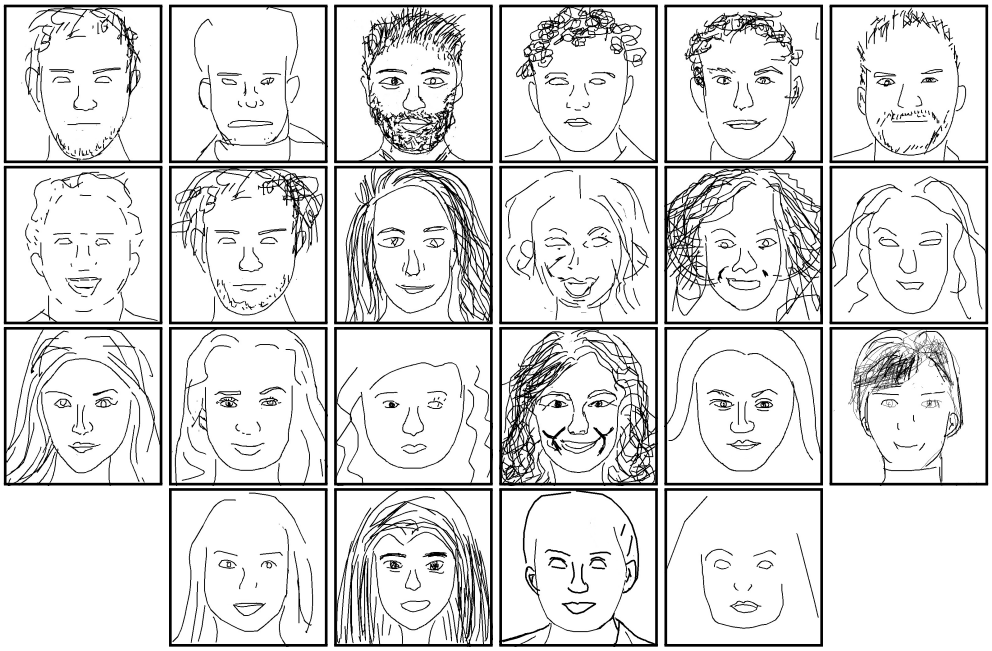

Fig. 4. Sketch inputs used in the perceptive evaluation.

Fig. 5. Generation results given test set sketches. For columns from left to right: pix2pix, Lines2FacePhoto, pix2pixHD, iSketchNFill and Ours.

7.3 Image Synthesis Architectures

Our Image Synthesis module adopts a GAN architecture utilizing a generator and a discriminator to generate real face images from the fused feature maps. The details of the generator are illustrated in Table 6. The discriminator employs a multi-scale discriminating manner: scale the input feature maps and the generated images in three different levels and go through three different sub-discriminators. See Table 7 for more details.

F(x)
Conv2d
BatchNorm2d
LeakyReLU

Table 2. **Conv2D-Block**

F(x)
ConvTranspose2d
BatchNorm2d
LeakyReLU

Table 3. **ConvTrans2D-Block**

Layer	Output Size	Filter
Input	$1 \times H \times W$	
Conv2D-Block	$32 \times \frac{H}{2} \times \frac{W}{2}$	$1 \rightarrow 32$
Resnet-Block	$32 \times \frac{H}{2} \times \frac{W}{2}$	$32 \rightarrow 32 \rightarrow 32$
Conv2D-Block	$64 \times \frac{H}{4} \times \frac{W}{4}$	$32 \rightarrow 64$
Resnet-Block	$64 \times \frac{H}{4} \times \frac{W}{4}$	$64 \rightarrow 64 \rightarrow 64$
Conv2D-Block	$128 \times \frac{H}{8} \times \frac{W}{8}$	$64 \rightarrow 128$
Resnet-Block	$128 \times \frac{H}{8} \times \frac{W}{8}$	$128 \rightarrow 128 \rightarrow 128$
Conv2D-Block	$256 \times \frac{H}{16} \times \frac{W}{16}$	$128 \rightarrow 256$
Resnet-Block	$256 \times \frac{H}{16} \times \frac{W}{16}$	$256 \rightarrow 256 \rightarrow 256$
Conv2D-Block	$512 \times \frac{H}{32} \times \frac{W}{32}$	$256 \rightarrow 512$
Resnet-Block	$512 \times \frac{H}{32} \times \frac{W}{32}$	$512 \rightarrow 512 \rightarrow 512$
Fully Connected	512	$512 \times \frac{H}{32} \times \frac{W}{32} \rightarrow 512$
Fully Connected	$512 \times \frac{H}{32} \times \frac{W}{32}$	$512 \rightarrow 512 \times \frac{H}{32} \times \frac{W}{32}$
Resnet-Block	$512 \times \frac{H}{32} \times \frac{W}{32}$	$512 \rightarrow 512 \rightarrow 512$
ConvTrans2D-Block	$256 \times \frac{H}{16} \times \frac{W}{16}$	$512 \rightarrow 256$
Resnet-Block	$256 \times \frac{H}{16} \times \frac{W}{16}$	$256 \rightarrow 256 \rightarrow 256$
ConvTrans2D-Block	$128 \times \frac{H}{8} \times \frac{W}{8}$	$256 \rightarrow 128$
Resnet-Block	$128 \times \frac{H}{8} \times \frac{W}{8}$	$128 \rightarrow 128 \rightarrow 128$
ConvTrans2D-Block	$64 \times \frac{H}{4} \times \frac{W}{4}$	$128 \rightarrow 64$
Resnet-Block	$64 \times \frac{H}{4} \times \frac{W}{4}$	$64 \rightarrow 64 \rightarrow 64$
ConvTrans2D-Block	$32 \times \frac{H}{2} \times \frac{W}{2}$	$64 \rightarrow 32$
Resnet-Block	$32 \times \frac{H}{2} \times \frac{W}{2}$	$32 \rightarrow 32 \rightarrow 32$
ConvTrans2D-Block	$32 \times H \times W$	$32 \rightarrow 32$
Resnet-Block	$32 \times H \times W$	$32 \rightarrow 32 \rightarrow 32$
ReflectionPad2d		
Conv2D-Block	$1 \times H \times W$	$32 \rightarrow 1$
Output	$1 \times H \times W$	

Table 4. The architecture of the Component Embedding Module.

Layer	Output Size	Filter
Input	512	
Fully Connected	$512 \times \frac{H}{32} \times \frac{W}{32}$	$512 \rightarrow 512 \times \frac{H}{32} \times \frac{W}{32}$
Resnet-Block	$512 \times \frac{H}{32} \times \frac{W}{32}$	$512 \rightarrow 512 \rightarrow 512$
ConvTrans2D-Block	$256 \times \frac{H}{16} \times \frac{W}{16}$	$512 \rightarrow 256$
Resnet-Block	$256 \times \frac{H}{16} \times \frac{W}{16}$	$256 \rightarrow 256 \rightarrow 256$
ConvTrans2D-Block	$256 \times \frac{H}{8} \times \frac{W}{8}$	$256 \rightarrow 256$
Resnet-Block	$256 \times \frac{H}{8} \times \frac{W}{8}$	$256 \rightarrow 256 \rightarrow 256$
ConvTrans2D-Block	$128 \times \frac{H}{4} \times \frac{W}{4}$	$256 \rightarrow 128$
Resnet-Block	$128 \times \frac{H}{4} \times \frac{W}{4}$	$128 \rightarrow 128 \rightarrow 128$
ConvTrans2D-Block	$64 \times \frac{H}{2} \times \frac{W}{2}$	$128 \rightarrow 64$
Resnet-Block	$64 \times \frac{H}{2} \times \frac{W}{2}$	$64 \rightarrow 64 \rightarrow 64$
ConvTrans2D-Block	$64 \times H \times W$	$64 \rightarrow 64$
Resnet-Block	$64 \times H \times W$	$64 \rightarrow 64 \rightarrow 64$
ReflectionPad2d		
Conv2D-Block	$32 \times H \times W$	$64 \rightarrow 32$
Output	$32 \times H \times W$	

Table 5. The architecture of the Feature Matching Module.

Generator		
Layer	**Output Size**	**Filter**
Input	$32 \times 512 \times 512$	
Conv2D-Block	$56 \times 512 \times 512$	$32 \rightarrow 56$
Conv2D-Block	$112 \times 256 \times 256$	$56 \rightarrow 112$
Conv2D-Block	$224 \times 128 \times 128$	$112 \rightarrow 224$
Conv2D-Block	$448 \times 64 \times 64$	$224 \rightarrow 448$
Resnet-Block ($\times 9$)	$448 \times 64 \times 64$	$448 \rightarrow 448 \rightarrow 448$
ConvTrans2D-Block	$224 \times 128 \times 128$	$448 \rightarrow 224$
ConvTrans2D-Block	$112 \times 256 \times 256$	$224 \rightarrow 112$
ConvTrans2D-Block	$224 \times 512 \times 512$	$448 \rightarrow 224$
ConvTrans2D-Block	$3 \times 512 \times 512$	$224 \rightarrow 3$
Output	$3 \times 512 \times 512$	

Table 6. The architecture of Generator in the Image Synthesis Module.

Discriminating Unit (DisUnit)		
Layer	**Output Size**	**Filter**
Input	$(32+3) \times H \times W$	
Conv2D-Block	$64 \times \frac{H}{2} \times \frac{W}{2}$	$(32+3) \rightarrow 64$
Conv2D-Block	$128 \times \frac{H}{4} \times \frac{W}{4}$	$64 \rightarrow 128$
Conv2D-Block	$256 \times \frac{H}{8} \times \frac{W}{8}$	$128 \rightarrow 256$
Conv2D-Block	$512 \times \frac{H}{8} \times \frac{W}{8}$	$256 \rightarrow 512$
Conv2D-Block	$512 \times \frac{H}{8} \times \frac{W}{8}$	$512 \rightarrow 512$
Output	$512 \times \frac{H}{8} \times \frac{W}{8}$	

Discriminator			
Layer	**D1 Output Size**	**D2 Output Size**	**D3 Output Size**
Input	$35 \times 512 \times 512$	$35 \times 512 \times 512$	$35 \times 512 \times 512$
AvgPool	-	$35 \times 256 \times 256$	$35 \times 256 \times 256$
AvgPool	-	-	$35 \times 128 \times 128$
DisUnit	$512 \times 64 \times 64$	$512 \times 32 \times 32$	$512 \times 32 \times 32$
Output	$512 \times 64 \times 64$	$512 \times 32 \times 32$	$512 \times 32 \times 32$

Table 7. The architecture of the Discriminator in Image Synthesis Module.