Generating Accurate Human Face Sketches from Text Descriptions

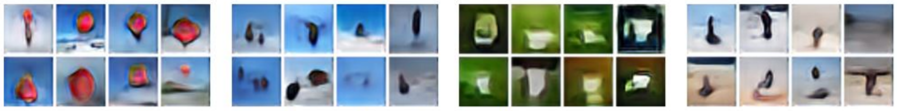
Jean-Michael Diei, Debdas Ghosh, Jeffrey Kim, Hyukjae Kwark
{jdiei,debdasg,jmkim1,hkwark}@andrew.cmu.edu

1 Introduction

Drawing a face for a suspect just based on the descriptions of the eyewitnesses is a difficult task. It requires professional skills and rich experience. It also requires a lot of time. However, with a well trained text-to-face model, anyone could directly generate photo-realistic faces of suspects based on the descriptions of eyewitnesses quickly. Most previous research on sketch generation assume that the original photo is available, which are usually unavailable from description of suspects. Since text-to-face synthesis is a sub-domain of text-to-image synthesis, there are only a few research focusing in this sub-domain (although it has more relative values in the public safety domain). There are some state-of-the-art methods in generating face images from text, but there is still a lot of room for improvement in similarity between input text and generated images. Here, we want to go even further and generate sketches rather than a RGB image.

2 Literature Review

We have found different papers regarding text to image conversion methods. One of them is Generating Images from Captions (https://arxiv.org/pdf/1511.02793.pdf), where they are taking textual descriptions as input and using them to generate relevant images. There are two components for this task, language modelling and image generation. There model alignDRAW (Align Deep Recurrent Attention Writer) uses Microsoft COCO dataset to accomplish these tasks. Below are the examples of their work.



A stop sign is flying in blue skies.

A herd of elephants flying in the blue skies.

A toilet seat sits open in the grass field.

A person skiing on sand clad vast desert.

Figure 1: Generating Images from Captions

Another paper Generative Adversarial Text to Image Synthesis (https://arxiv.org/pdf/1605.05396.pdf) uses Caltech-UCSD birds database and Oxford 102 Flowers dataset to generate plausible images of birds and flowers from detailed text descriptions. They are using DC GAN to counter these two subproblems, learn a text feature representation that captures the important visual details and use these features to synthesize a compelling image that a human might mistake for real. Examples of their work have shown in Figure 2.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

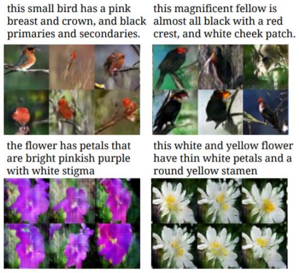

this small bird has a pink breast and crown, and black primaries and secondaries.
this magnificent fellow is almost all black with a red crest, and white cheek patch.
the flower has petals that are bright pinkish purple with white stigma
this white and yellow flower have thin white petals and a round yellow stamen

Figure 2: Generative Adversarial Text to Image Synthesis

3 Baseline Model

There are two major components of our model: *data generation* and *text-to-faceSketch generation*.

3.1 Data Generation

As of current, there are no pre-existing datasets that pair a text caption to a face sketch. Thus, our ultimate goal in this component of our model is to generate (text, faceSketch) dataset pairs to provide supervised learning in the *text-to-faceSketch generation* component. We plan to generate such dataset in the following procedure: First, generate text-to-faceImage using a AttnGAN, then generate faceImage-to-faceSketch using VAE GAN, and finally simply pair text with its corresponding faceSketch.

3.2 Text-to-FaceSketch Generation

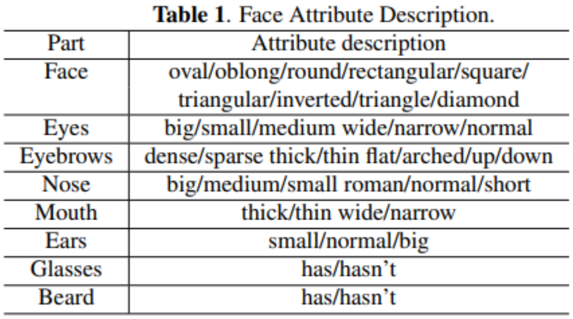


Table 1. Face Attribute Description.

Part	Attribute description
Face	oval/oblong/round/rectangular/square/ triangular/inverted/triangle/diamond
Eyes	big/small/medium wide/narrow/normal
Eyebrows	dense/sparse thick/thin flat/arched/up/down
Nose	big/medium/small roman/normal/short
Mouth	thick/thin wide/narrow
Ears	small/normal/big
Glasses	has/hasn't
Beard	has/hasn't

Figure 3: Text2Sketch Dataset attributes

We will be using the baseline model introduced in *Y. Wang et al. "Text2Sketch: Learning Face Sketch from Facial Attribute Text"* [3], which uses a Stagewise-GAN composed of two stages, to generate a high resolution (256×256) sketch. The pre-processed data (Text2Sketch dataset) is as the following figure above which was generated based on the categories used for criminal appearance study. The final dataset is a text-sketched image pair, where the sketch imaged is a result of post-processing with a state-of-the-art supervised descent method (SDM) for alignment. The text description is vectorized by frequency by building a dictionary of attributes, followed by a normalization.

The network model is composed of two stages where each stage has a Generator and Discriminator as shown in Figure 4 below. Stage 1 synthesizes the batch of low-resolution face sketches with a Gaussian noise. In the Generator, the vector representation of text description concatenated with a Gaussian noise matrix are passed through deconvolutional layers to be upsampled to a sketch of 64×64. The Discriminator downsamples by a stack of convolutional layers and then concatenated with the vectorized text. A convolutional layer with kernel size 4×4 followed by a Sigmoid is applied to get the classification score of the descriminator.[3] Stage II inputs a low resolution sketches and text descriptions to the second GAN to produce high resolution sketches(256×256), where the Gaussian noise is replaced with the low resolution sketches generated in stage I. The output of Stage II Discriminator is followed by a fully connected layer to produce a decision score.[3]

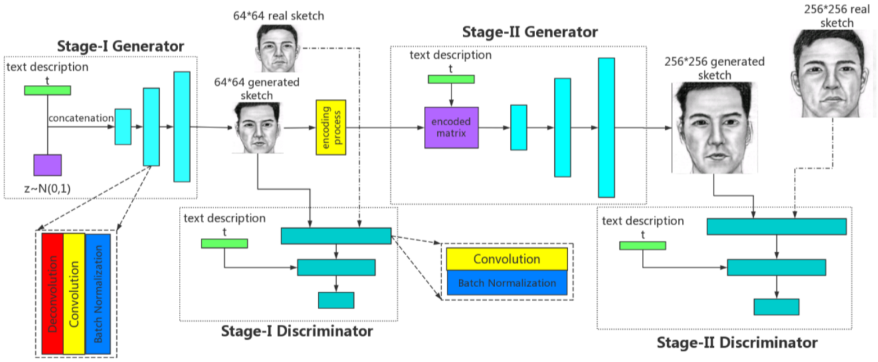


Figure 4: Text2Sketch network architecture

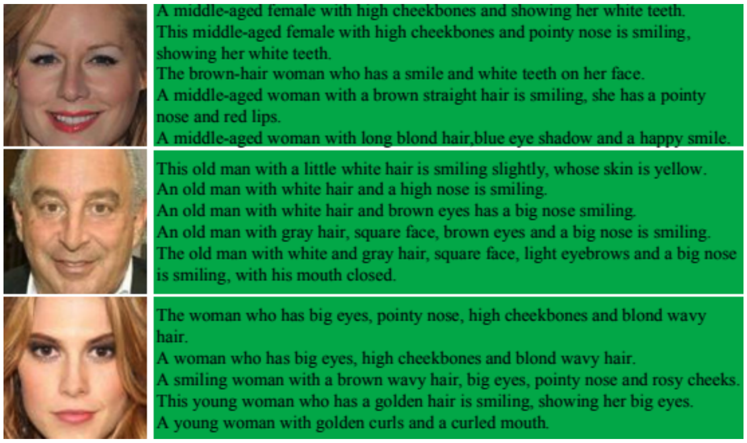
Stage I was trained with learning rate = 0.0001, decay ratio of 0.8 every 50 epochs, and Stage II was trained with learning rate = 0.0005, decay ratio of 0.8 every 50 epochs. Both stages were trained through 500 epochs with batch size of 36. Generators and discriminators were trained alternatively.[3]

Evaluation was performed through the following steps. First, you train a CNN model on a classification task for feature extraction of the CUFSF sketched images. Then compute the Jaccard distance of the result after running the output sketch to the CNN model, and see K nearest neighbors. If the true output exists among the K nearest neighbors, the output is marked to be correct.

4 Dataset

4.1 Text-FaceImage Pairs: *SCU-Text2face dataset*

This dataset is based on CelebA dataset, which contains 1000 images. For each face image, there are five text descriptions from 5 different people. This dataset could help build the baseline for text-to-faceImage synthesis task [1]. Some examples of SCU-Text2face dataset is shown in Figure 5.


A middle-aged female with high cheekbones and showing her white teeth.
This middle-aged female with high cheekbones and pointy nose is smiling, showing her white teeth.
The brown-hair woman who has a smile and white teeth on her face.
A middle-aged woman with a brown straight hair is smiling, she has a pointy nose and red lips.
A middle-aged woman with long blond hair,blue eye shadow and a happy smile.

This old man with a little white hair is smiling slightly, whose skin is yellow.
An old man with white hair and a high nose is smiling.
An old man with white hair and brown eyes has a big nose smiling.
An old man with gray hair, square face, brown eyes and a big nose is smiling.
The old man with white and gray hair, square face, light eyebrows and a big nose is smiling, with his mouth closed.

The woman who has big eyes, pointy nose, high cheekbones and blond wavy hair.
A woman who has big eyes, high cheekbones and blond wavy hair.
A smiling woman with a brown wavy hair, big eyes, pointy nose and rosy cheeks.
This young woman who has a golden hair is smiling, showing her big eyes.
A young woman with golden curls and a curled mouth.

Figure 5: Examples of SCU-Text2face dataset

4.2 FaceImage-FaceSketch Pairs: *CUFS dataset, CUFSF dataset*

The CUFS dataset (consisting of the CUHK student dataset [4], the AR dataset [5], and the XM2VTS dataset [6]) contains 606 faces and the CUFSF dataset contains 1194 faces. Some of the examples of CUFS dataset is shown in Figure 6. The CUFSF dataset [2] is more challenging than the CUFS dataset because (1) the photos were captured under different lighting conditions and (2) the sketches were made with shape exaggeration drawn by an artist when viewing the photos. Some of the examples of CUFS dataset is shown in Figure 7.

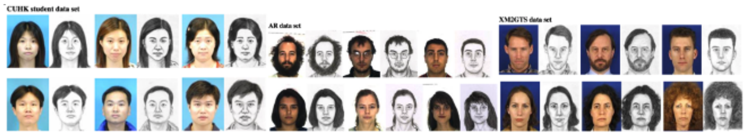


Figure 6: Examples of CUFS dataset

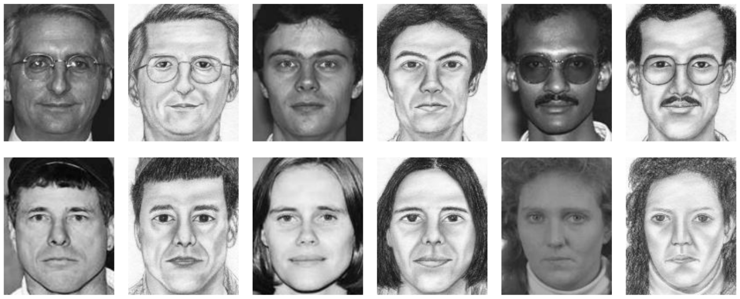


Figure 7: Examples of CUFSF dataset

Preprocessing In order to account for the misalignment and the variational lighting in the images, we can develop a process to clean these images. For alignment, we can warp the images so that the eyes in the sketches are at specific chosen coordinates. For lighting, we can leverage OpenCV to equalize contrast and brightness of the sketches.

References

[1] Chen, X., Qing, L., He, X., Luo, X., Xu, Y.: FTGAN: A fully-trained generative adversarial networks for text to face generation. CoRR abs/1904.05729 (2019)

[2] W. Zhang, X. Wang and X. Tang. Coupled Information-Theoretic Encoding for Face Photo-Sketch Recognition. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[3] Y. Wang et al., "Text2Sketch: Learning Face Sketch from Facial Attribute Text," 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, 2018, pp. 669-673, doi: 10.1109/ICIP.2018.8451236.

[4] X. Wang and X. Tang, Face Photo-Sketch Synthesis and Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, Vol. 31, 2009.

[5] A. M. Martinez, and R. Benavente, "The AR Face Database," *CVC Technical Report 24*, June 1998.

[6] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: the Extended of M2VTS Database," *in Proceedings of International Conference on Audio- and Video-Based Person Authentication*, pp. 72-77, 1999.

[7] Han Zhang and Dimitris Metaxas "StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks", Aug 2017