International Journal of
Advanced Science Computing and Engineering

Journal Homepage www.ijasce.org/index.php/ijasce

Generating Accurate Human Face Sketches from Text Descriptions

Shorya Sharma [a,*]

[a] School of Electrical Sciences, Indian Institute of Technology Bhubaneswar, Bhubaneswar, Odisha, 751013, India
Corresponding author: *ss118@iitbbs.ac.in

Abstract— **Drawing a face for a suspect just based on the descriptions of the eyewitnesses is a difficult task. There are some state-of-the-art methods in generating images from text, but there are only a few research in generating face images from text and close to none in generating sketches from text. As a result, there is no dataset available to tackle this task. In this paper, we generated a new text-to-sketch dataset for our novel task, and provide two attention based SOTA GAN end-to-end models, Attn_LSTM_256 and Attn_GRU_512, trained on the dataset resulting in Inception score of 1.868 and 1.902, and FID of 175.46 and 176.98. We further propose possible future improvements by applying different model architectures or preserving performance with simplified architectures for real-world applications.**

Keywords— **GAN;LSTM;GRU.**

Manuscript received 20 Feb. 2024; revised 25 Mar. 2024; accepted 10 Apr. 2024. Date of publication 30 Apr. 2024.

I. Introduction

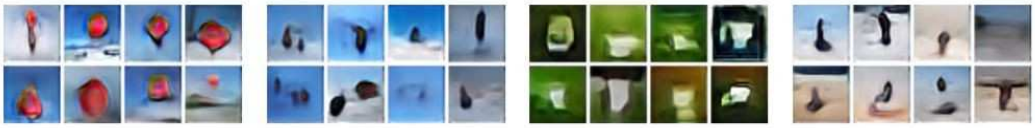
Drawing a face for a suspect just based on the descriptions of the eyewitnesses is a challenging task. It requires professional skills and rich experience. It also requires a lot of time. However, even with a well trained text-to-face model, it will not be able to directly generate photo-realistic faces of suspects based on the descriptions of eyewitnesses. They either produce unclear images or generate images at a low resolution, making the images impossible to use for our identification task. This is due to the fact that photo-realism is simply too hard to capture from text at this moment, thus we decided to constrain ourselves to sketches as we do not need to generate as detailed a face in order to see good results. Most previous research on sketch generation of a face assume that the original photo is available, which are usually unavailable from description of suspects. Also, 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). This is also mainly due to lack of public (text description, face image/sketch) dataset. There are some state-of-the-art methods in generating bird/flower images from text, but there are only a few literature in generating a face image from a text and close to none in generating a face sketch from a text description. Here, not only have we tackled a novel task, but we have also generated a new dataset of text description and face sketch pairs.

II. Material and Method

We have found different papers regarding text to image conversion methods. One of them is Generating Images from Captions[16], 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. Their model, alignDRAW (Align Deep Recurrent Attention Writer), uses Microsoft COCO dataset to accomplish these tasks. Figure 1 is the examples of their work.

Another paper, Generative Adversarial Text to Image Synthesis[15] 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.

An attempt to generate face images from text descriptions can be seen in "Text2FaceGAN: Face Generation from Fine Grained Textual Descriptions" [13]. This paper leverages the CelebA dataset and generates a text to face dataset by auto-generating captions given the attributes of each image. These

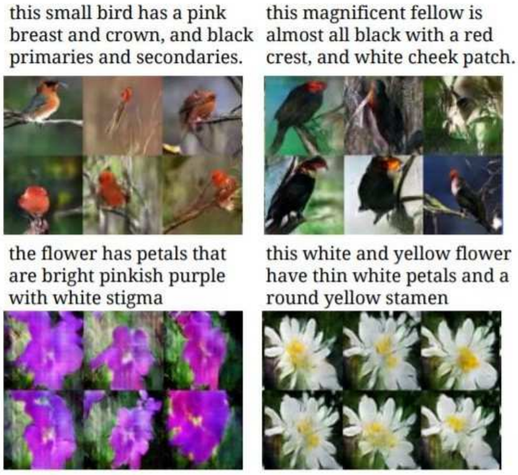

A stop sign is flying in blue skies.

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.

Fig. 1 Generating Images from Captions


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

Fig. 2 Generative Adversarial Text to Image Synthesis

captions include descriptions of six sentences based on structure of the face, facial hair, hairstyle, fine grain face details, accessories worn.[13] The model builds upon Generative Adversarial Networks (GAN) and Matching-aware Discriminator (GAN-CLS) for face synthesis from text embeddings. A pre-trained model of skip-thought Vectors was used to encode the input text which are known to obtain very good results for image retrieval task.[13] A representative output presented in the paper can be seen in Figure 3. Although producing faces that align to the attributes in the descriptions, the paper generates low resolution face images, which are hard to use for identification.

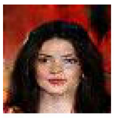

The woman has high cheekbones. She has straight hair which is black in colour. She has big lips with arched eyebrows. The smiling, young woman has rosy cheeks and heavy makeup. She is wearing lipstick.

Fig. 3 Face Generation from Fine Grained Textual Descriptions

To improve on the quality of image synthesis from text descriptions, the state of the art was introduced in Tao Xu and Xiaolei Huang’s paper "AttnGan: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks". The model includes a layered attentional GAN which is able to automatically select the condition at the word level for generating different parts of the image.[12] AttnGAN achieves a high resolution image by combining the regional image vectors, generating new image features for each sub-regions. AttnGAN also introduces a Deep Attentional Multimodal Similarity Model (DAMSM), which computes the similarity between the generated image and the sentence. Figure 4 is an example of the result of AttnGAN.

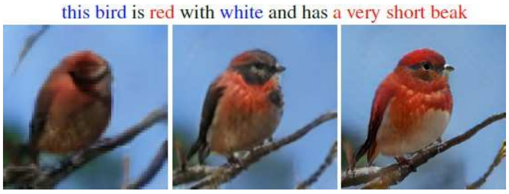


Fig. 4 Low-to-High resolution images generated by AttentionGAN

Recently, a Text-to-Face Generation via Attribute Disentanglement was researched from University of Queensland.[8] This paper uses a CelebA dataset and StyleGAN to achieve a high resolution of a set of images with diversity based on text descriptions. The ultimate goal was to provide the witness with a set of generated faces based on the description, and ask the witness to pick among the diverse set of outputs which one resembles the suspect most. The model uses a multi-label classifer (T) that generates 40 facial attributes from free form natural language descriptions, and a image encoder (E) from MobileNet to obtain image embeddings, and uses a Pre-trained model of Style-GAN2 to generate the final set of images. The paper utilizes a noise vector as an input to ensure diversity. Additionally, the paper proposes 4 noise vector manipulation techniques (differentiation, nonlinear re-weighting, normalization, and feature lock) to improve performance of the model [8]. The literature introduced above provide solutions to the task of text to image synthesis through varying structures such as Stack-GAN, Conditional-GAN, DC-GAN, Style-GAN, and AttnGAN. However, there is no preliminary work focusing on text to face sketch image synthesis which is a common task for making facial composites of suspects.

A. Contribution

We decided to leverage a preexisting architecture to solve our text-to-sketch task. Due to the nature of our problem, we knew we needed two encoders, one for text and one for the image. In order to perform Sketch generation from text, it is important to sample from the distribution of our sketch images based upon the most critical words from the text. Because of this, we pursued AttnGAN as it "allows attention-driven, multi-stage refinement for fine-grained text-to-image generation."[12] AttnGAN consists of two important components: the attentional generative network and the deep attentional multimodal similarity model. The attention model enables the generative network to select particular parts of the image based on words that are most relevant to those sub-regions. The DAMSM learns two neural networks that map part of an image and words from the sentence into the same subspace, which allows for computing image-text similarity at the word level to calculate a loss for image generation. We leveraged this work by applying and training this AttnGAN

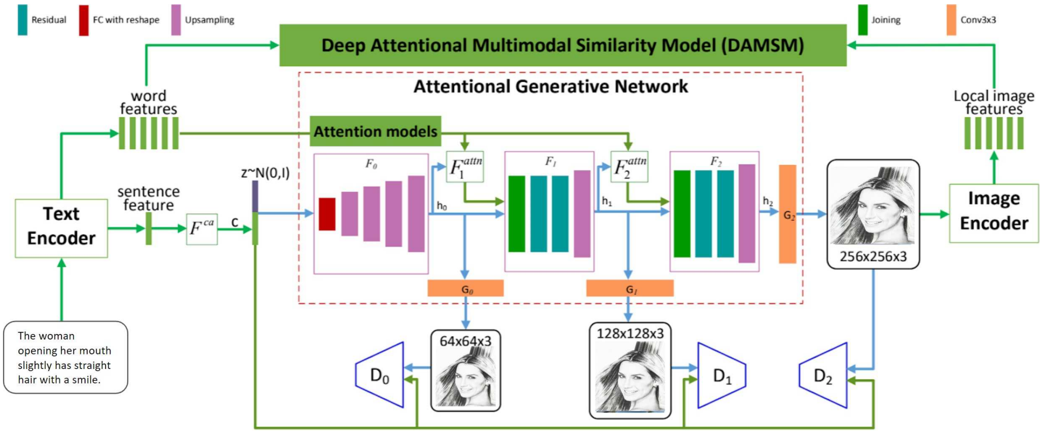
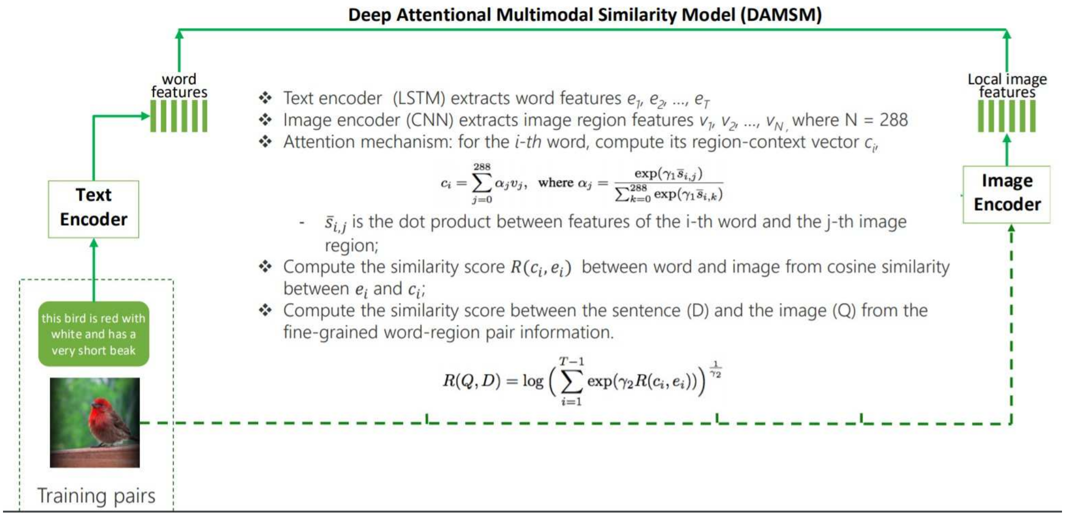


Fig. 5 Adapted AttentionGAN architecture


Deep Attentional Multimodal Similarity Model (DAMSM)
word features
Local image features
Text Encoder
Image Encoder
this bird is red with white and has a very short beak
Training pairs

- ❖ Text encoder (LSTM) extracts word features $e_1, e_2, \ldots, e_T$
- ❖ Image encoder (CNN) extracts image region features $v_1, v_2, \ldots, v_N$, where N = 288
- ❖ Attention mechanism: for the *i-th* word, compute its region-context vector c_i,

$$c_i = \sum_{j=0}^{288} \alpha_j v_j, \quad \text{where } \alpha_j = \frac{\exp(\gamma_1 \bar{s}_{i,j})}{\sum_{k=0}^{288} \exp(\gamma_1 \bar{s}_{i,k})}$$

  - $\bar{s}_{i,j}$ is the dot product between features of the i-th word and the j-th image region;
- ❖ Compute the similarity score $R(c_i, e_i)$ between word and image from cosine similarity between e_i and c_i;
- ❖ Compute the similarity score between the sentence (D) and the image (Q) from the fine-grained word-region pair information.

$$R(Q, D) = \log\Big(\sum_{i=1}^{T-1} \exp(\gamma_2 R(c_i, e_i))\Big)^{\frac{1}{\gamma_2}}$$

Fig. 6 Deep Attentional Multimodal Similarity Model Architecture

using our new dataset in order to solve the text-to-Sketch problem.

B. Model Details

The AttnGan consists of the Attentional Generative Network and Deep Attentional Multimodal Similarity Model.The text gets encoded and is used within the attention models and the DAMSM. The Attentional Generative Network (which is within the dashed red box in Figure 5) contains the attention models that enables the generative networks to extract particular parts of the image conditioned on relevant words. The output of the highest resolution Generator (G2) is passed through an image encoder, which is built upon a pretrained Inception-v3 model.The DAMSM, as shown in more details in Figure 6, utilizes features from the Text Encoder and the Image Encoder, we can then compute image-text similarity at the word level.

To further clarify, the DAMSM consists of two neural networks, the image encoder (RNN) and the Image Encoder (CNN) and uses the word features and intermediate feature maps to compute the similarity.

C. Experiment

We have trained our Deep Attentional Multimodal Similarity Model (DAMSM) with different RNN types like Long short-term memory(LSTM) and Gated Recurrent Unit(GRU). We have also tried different embedding layer sizes and regularization methods like gradient clipping, noise addition etc. While training DAMSM is faster, which takes about 20 minutes per epoch, training AttnGAN is quite time consuming as it takes more than 3 hours per epoch. To get a reasonable result it is required to train for at least 30 epochs. Given this time constraint we were limited with our experiments even with multiple AWS instances. Based on our experiments we have found that with LSTM GRU type, 512 embedding dimension, 0.25 gradient clipping, 0.5 dropout works best for DAMSM. For generator we have used GLU activation, instead of ReLU and for the discriminator we have used LeakyReLU. We experimented on two optimization methods SGD and Adam, and figured that Adam works better than SGD.

D. Dataset

Currently, there are no public text and face sketch dataset available. Thus, we have created our own text and face sketch pair dataset based upon the CelebA dataset. Our generated dataset will have the same number of samples as CelebA dataset (http://mmlab.ie.cuhk.edu.hk/projects/ CelebA.html), which includes 10,177 identities and 202,599 face images.

E. Text Generator

We have used the auto-generated text descriptions from the 40 features in CelebA from https: //github.com/2KangHo/AttnGAN-CelebA. The 40 boolean features of CelebA dataset are ’5 o Clock Shadow’, ’Arched Eyebrows’, ’Attractive’, ’Bags Under Eyes’, ’Bald’, ’Bangs’, ’Big Lips’, ’Big Nose’, ’Black Hair’, ’Blond Hair’, ’Blurry’, ’Brown Hair’, ’Bushy Eyebrows’, ’Chubby’, ’Double Chin’, ’Eyeglasses’, ’Goatee’, ’Gray Hair’, ’Heavy Makeup’, ’High Cheekbones’, ’Male’, ’Mouth Slightly Open’, ’Mustache’, ’Narrow Eyes’, ’No Beard’, ’Oval Face’, ’Pale Skin’, ’Pointy Nose’, ’Receding Hairline’, ’Rosy Cheeks’, ’Sideburns’, ’Smiling’, ’Straight Hair’, ’Wavy Hair’, ’Wearing Earrings’, ’Wearing Hat’, ’Wearing Lipstick’, ’Wearing Necklace’, ’Wearing Necktie’, ’Young’. These auto-generated texts are generated by randomly selecting binary features in the CelebA and putting them into a contextual description. For example, if "Wavy Hair" feature is 1 and "Male" feature is 0, the auto-generated text could be: "The woman has wavy hair" or "A woman’s hair is wavy" with the articles, "A" and "The" generated randomly as well.

F. Skecth Generation

To generate sketch images from the RGB images in the CelebA dataset, we have trained a CycleGAN using the CUFS and 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. 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.

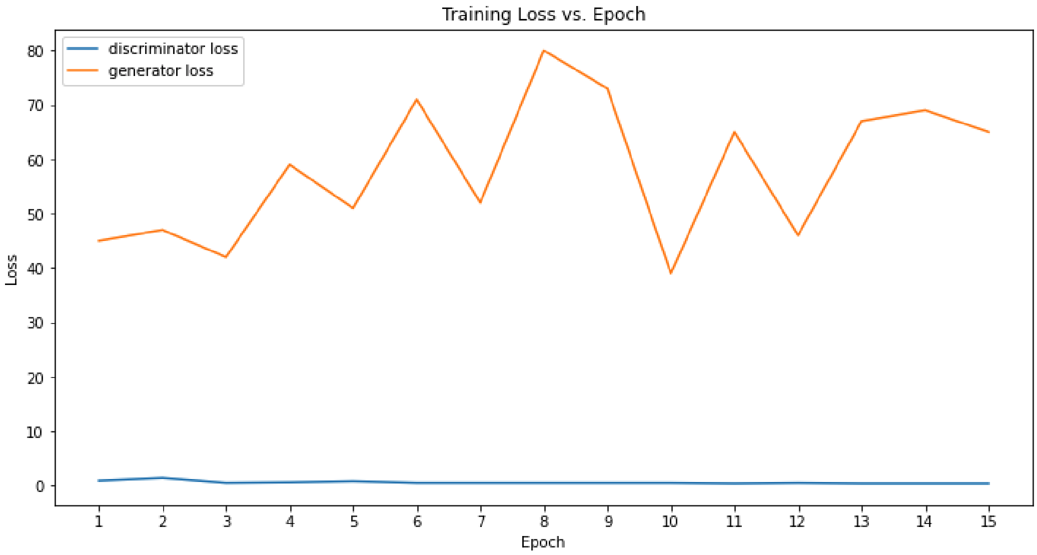
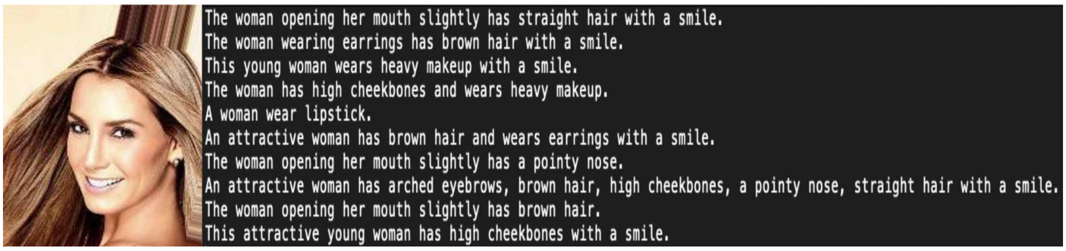


Fig. 7 Final Model Training



```
The woman opening her mouth slightly has straight hair with a smile.
The woman wearing earrings has brown hair with a smile.
This young woman wears heavy makeup with a smile.
The woman has high cheekbones and wears heavy makeup.
A woman wear lipstick.
An attractive woman has brown hair and wears earrings with a smile.
The woman opening her mouth slightly has a pointy nose.
An attractive woman has arched eyebrows, brown hair, high cheekbones, a pointy nose, straight hair with a smile.
The woman opening her mouth slightly has brown hair.
This attractive young woman has high cheekbones with a smile.
```

Fig. 8 Auto-generated Text Descriptions from CelebA Dataset

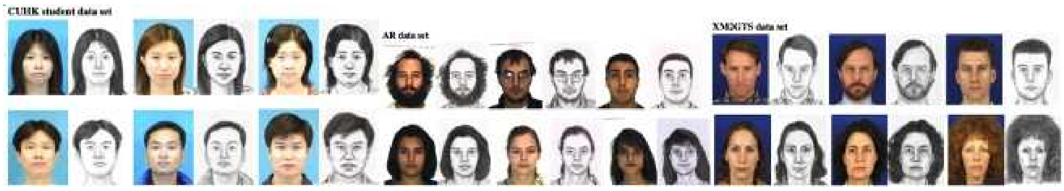


Fig. 9 Examples of CUFS dataset

G. Baseline

As there is no pre-existing dataset to train any model, generating the dataset itself is the baseline. The text is auto-generated from the 40 binary features in the CelebA dataset. For a generated text, we generate a sketch from the corresponding RGB CelebA image using the CycleGAN. To train the CycleGAN, we utilized the images from the Celebrity dataset along with the sketches from CUFSF dataset as depicted in Figure 9 and Figure 10. As this is an unpaired

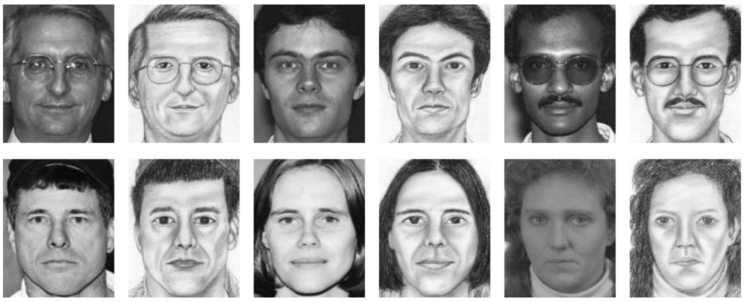


Fig. 10 Examples of CUFSF dataset

training process, there is no need to have matching Face-Sketch pairs. This allows us to use any Face dataset in combination with a Sketch dataset in order to produce a decent result. An example output of the CycleGAN during its training procedure is depicted in Figure 11. Like we mentioned previously, the caveat for GANs are difficult to obtain good results from, and because of this, it is challenging to measure performance. A full diagram of our baseline is depicted in Figure 12.

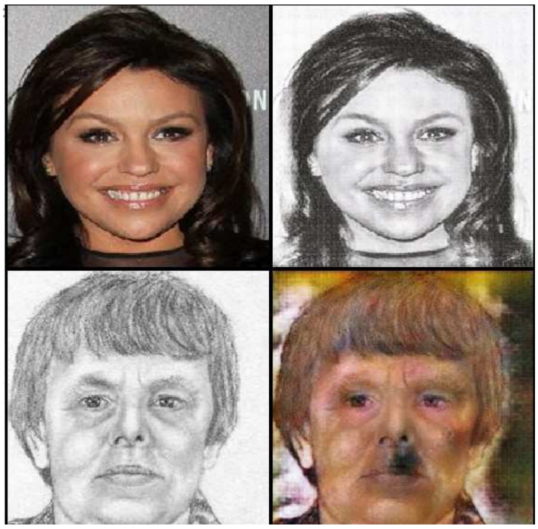


Fig.11 Top Left: Real Face, Bottom Left; Real Sketch, Top Right: Fake Sketch, Bottom Right:Fake Face

H. Evaluation METRIC

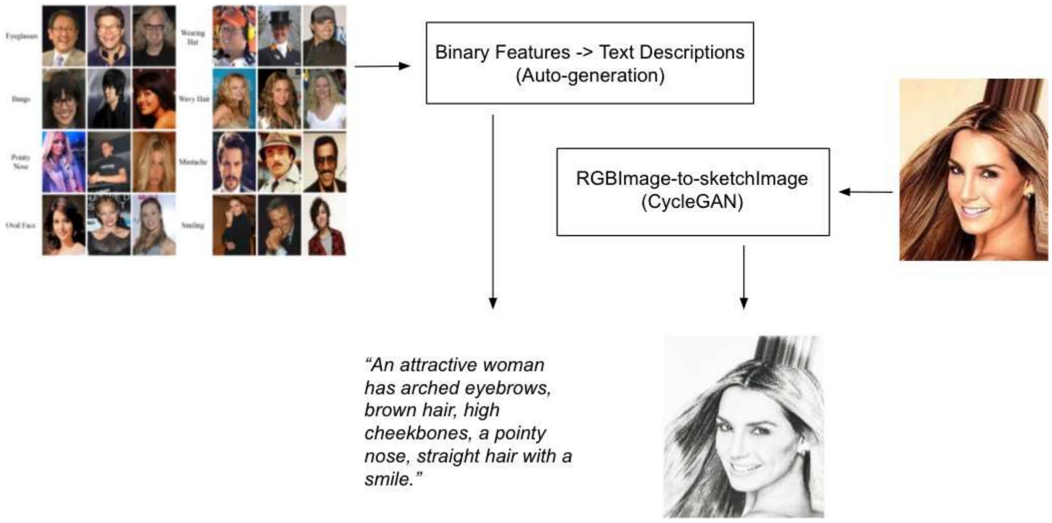
Two commonly used evaluation metrics for the images generated through GAN-like structures are Inception score, Fréchet Inception distance (FID). Inception score is known to measure the diversity of generated images. We expect Inception score to help find models that generate a more diverse set of faces that match the text descriptions. FID is inversely correlated with Inception score, which will help find models produce a similar synthesized data as the real distribution. Since Inception score and FID are the most popular evaluation metrics, we will use them to compare the results of other models based on the CelebA dataset to compare our model's quality of generated sketches.

I. Inception Score

Inception score is one of the most common ways to evaluate GAN. Inception score measures the quality of a generated image by computing the KLdivergence between the (logit) response produced by this image and the marginal distribution, i.e., the average response of all the generated images, using an Inception network trained on ImageNet.[11] This score focuses on the diversity of the images rather than the real image. The score is given as the following equation.[11]

$$IS(G) = exp(E_{x\sim p_g}[D_{KL}(p(y|x)||p(y))]) \qquad (1)$$

Where x is a generated sample from the learned generator ditribution p_g and D_{KL} is the KL divergence between the conditional class distribution $(p(y|x)$ and marginal class



Eyeglasses
Bangs
Pointy Nose
Oval Face
Wearing Hat
Wavy Hair
Mustache
Smiling
Binary Features -> Text Descriptions (Auto-generation)
RGBImage-to-sketchImage (CycleGAN)
"An attractive woman has arched eyebrows, brown hair, high cheekbones, a pointy nose, straight hair with a smile."

Fig. 12 Baseline/Dataset Generation

distribution $p(y) = E_{x \sim p_g}[p(y|x)]$. [11] y is the label according to the inception network.

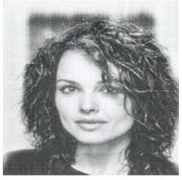
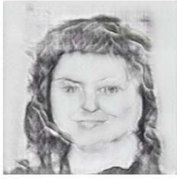
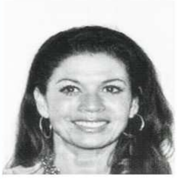
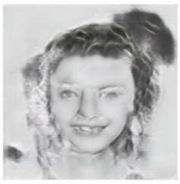
Original Image	Generated Image	Text Description
		The woman has bangs, big lips, high cheekbones, an oval face, a pointy nose, wavy hair with a smile. The attractive woman has an oval face and wears heavy makeup
		This old woman puts on lipstick with a smile. The woman opening her mouth slightly has heavy makeup. The old woman wears earrings. A woman has black hair with a smile.

Fig. 13 Sample result of sketch generation from text description from test dataset.

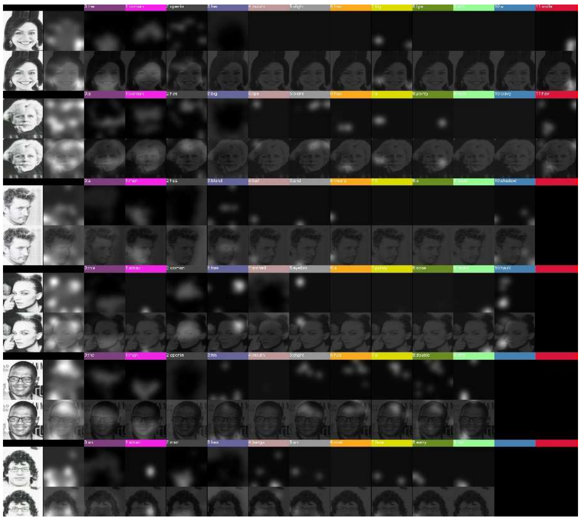


Fig. 14 Attention Map of DAMSM after 100 epochs

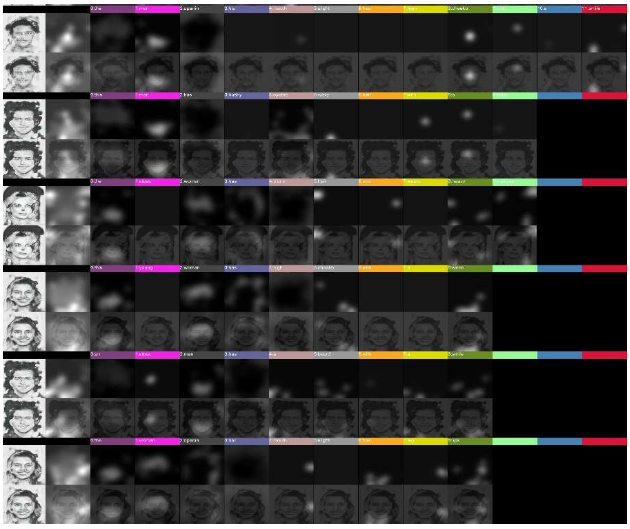


Fig. 15 Attention Map of AttnGAN after 10 epochs

J. *Fréchet Inception Distance (FID)*

Fréchet Inception distance (FID) compares Inception embeddings' distribution (responses of the penultimate layer of the Inception network) of the real and generated images.[11] The distance (FID) is defined as the following equation.

$$d^2\left((m_r, C_r), (m_g, C_g)\right) = ||m_r - m_g||^2 + T_r(C_r + C_g - 2(C_r C_g)^{1/2}) \quad (2)$$

where (m_r, C_r) and (m_g, C_g) is the mean and covariance of the real image and generated image respectively. This comparison approximates the activations of real and generated images as Gaussian distributions, computing their means and covariances.

In general, Inception score estimates the quality and diversity of the collection of the gener- ated images through the inception V3 model. For the application to this task, Inception score would be a good metric to evaluate the quality of synthesized images, but could be harsh on grading the diversity of the images since we only generate faces. FID score also uses the inception V3 model, but it extracts the CV-specific features of the input image and evaluates the closeness of generated data distribution and the real data distribution. FID can be a good metric to evaluate the performance of the model for this task, but may lack the ability to better evaluate models that generate a diverse set of sketch face images which is preferred in the real-world scenario. Based on its pros and cons, we included both metrics for evaluating our task of generating sketch images.

III. Results and Discussion

A. *Generated Image*

A representative example of text to sketch generation from our trained AttnGAN can be found in Figure 13. Note that these examples were sampled from the test dataset, which were never introduced to the model. The original image is also included in the leftmost column for reference. The figure shows that the model is capable of capturing facial characteristics such as "wavy hair", "high cheekbones", "opening mouth", "woman", and "smile".

B. Evaluation

Through experiments, we have achieved two models, AttnGAN with LSTM RNN, embedding size 256 (Attn_LSTM_256), and AttnGAN with GRU RNN, embedding size 512 (Attn_GRU_512) as our best performing SOTA. For evaluation of the trained generator, two evaluation metrics, Inception score and Fréchet Inception Distance(FID) were measured. The most widely used inception network, Inception V3, was used for evaluation of both metrics. The two metrics were calculated with 200 sample images from an unseen test dataset of text-sketch pairs. The model with the lowest generator KL loss value were selected for both models, Attn_LSTM_256 and Attn_GRU_512, which were epoch 34 and epoch 15 respectively. The evaluation results of both models are shown in Table 1. As a result, we were able to achieve a state-of-the-art models Attn_LSTM_256 and Attn_GRU_512.

TABLE I
EVALUATION OF ATTENTION MODELS

Model	Inception score	FID
Attn_LSTM_256	1.868 ± 0.196	175.46
Attn_GRU_512	1.902 ± 0.189	176.98

C. Attention Map

We have plotted attention map to observe which words are responsible of generating each section of the image. The corresponding region responsible for each text is marked white on the generated image. Figure 14 is a sample attention map that we got after training the encoder for 100 epochs. It can be seen that it is focusing on lips for smiles and eyes for detecting face structure of man and woman.

With more epochs the results are more prominent. After training the AttnGAN for 10 epochs, we can see in Figure 15 that the attention map successfully evolved to capture facial characteristics such as nose, eyes and lips. We could observe that the generated images correctly capture the facial characteristics mentioned in the text descriptions from attention maps.

IV. CONCLUSION

In this paper, we successfully created a text to sketch dataset based on the CelebA dataset containing 200,000 celebrity images, which can be further utilized to exploit the novel task of generating police sketches from text descriptions. Furthermore, we have proven that application of AttnGAN to generate sketch images successfully capture the facial characteristics included in text descriptions. We have also found the best configuration of AttnGAN and its variant by experimenting on different RNN types and embedding sizes. Additionally, we have provided the most widely used metric values (Inception score, FID) for the two-attention based SOTA model we have achieved. However, we have found room for improvement during the application of the model. Through experiments on a new dataset containing 200 sketch images provided by Beijing Normal University, we have found that the model is prone to failing on descriptions that include long sentences or unseen words. Failing to capture characteristics from such text descriptions result into low diversity and unrealistic images, having great impact on the models overall performance. For future improvements, we recommend trying different model architectures such as Stack-GAN, Conditional-GAN, DC-GAN, and Style-GAN which are known to be powerful in solving face image generation tasks. Furthermore, simplifying the model architecture while preserving its performance would be another direction for future research such that it could be deployable to various mobile devices for real-world applications.

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), Oct. 2018, 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] H. Zhang and D. Metaxas "StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks", Aug 2017

[8] T. Wang, T. Zhang, and B. C. Lovell, "Faces la Carte: Text-to-Face Generation via Attribute Disentanblement", Sep 2020

[9] Z. Liu, P. Luo, X. Wang, X. Tang: Deep Learning Face Attributes in the Wild. Proceedings of International Conference on Computer Vision (ICCV), (2015)

[10] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks,” 2017 IEEE International Conference on Computer Vision (ICCV), Oct. 2017, doi: 10.1109/iccv.2017.244.

[11] K. Shmelkov, C. Schmid, and K. A. Inria, "How good is my GAN?", ECCV (2018)

[12] Tao Xu, Xiaolei Huang, "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks" 1711.10485v1/cs.CV, 28 Nov 2017

[13] O. R. Nasir, S. K. Jha, and M. S. Grover, "Text2FaceGAN: Face Generation from Fine Grained Textual Descriptions", 1911.11378v1/cs.LG, 26 Nov 2019

[14] IIIT-D Sketch Database: http://www.iab-rubric.org/resources/sketchDatabase.html

[15] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. arXiv preprint arXiv:1605.05396, 2016.

[16] E. Mansimov, E. Parisotto, J. Lei Ba, and R. Salakhutdinov "Generating Images From CaptionsWith Attention" 1511.02793v2/cs.LG, 29 Feb 2016.