International Journal of
Advanced Science Computing and Engineering

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

Generating Accurate Human Face Sketches from Text Descriptions

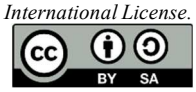
Shorya Sharma [a,*]

[a] *School of Electrical Sciences, Indian Institute of Technology Bhubaneswar, Bhubaneswar, Odisha, 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 is only a little research in generating face images from text, and almost none in generating sketches from text. As a result, there is no dataset available to tackle this task. We developed a text-to-sketch dataset derived from the CelebA dataset, which comprises 200,000 celebrity images, thereby facilitating the investigation of the novel task of generating police sketches from textual descriptions. Furthermore, we demonstrated that the application of AttnGAN for generating sketch images effectively captures the facial features described in the text. We identified the optimal configuration for AttnGAN and its variants through experiments involving various recurrent neural network types and embedding sizes. We provided commonly used metric values, such as the Inception score and Fréchet Inception Distance (FID), for the two-attention-based state-of-the-art model we achieved. However, we also identified areas for improvement in the model's application. Experiments conducted with a new dataset consisting of 200 sketch images from Beijing Normal University revealed that the model encounters challenges when handling longer sentences or unfamiliar terms within descriptions. This limitation in capturing features from such text contributes to a decrease in image diversity and realism, adversely impacting the overall performance of the model. For future improvements, consider exploring alternative models such as Stack-GAN, Conditional-GAN, DC-GAN, and Style-GAN, which are known for their capabilities in face image generation. Simplifying architecture while maintaining performance can also help deploy models on mobile devices for real-world use.**

***Keywords*— Text-to-Sketch generation; facial composite synthesis; AttnGAN; Fréchet inception distance.**

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 based solely on eyewitness descriptions is a challenging task. It requires professional skills and rich experience. It also requires a significant amount 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 [1]. This is because photo-realism is currently too challenging to capture from text; thus, we decided to limit ourselves to sketches, as we do not need to generate as detailed a face to achieve good results [2]. Most previous research on sketch generation of a face assumes that the original photo is available, which is usually unavailable from a description of suspects. Also, since text-to-face synthesis is a sub-domain of text-to-image synthesis, only a few studies are focusing on this sub-domain (although it has more relevant values in the public safety domain). This is also mainly due to the lack of a public dataset (text description, face image/sketch) [3]. There are some state-of-the-art methods for generating bird/flower images from text. Still, there is only a little literature on generating a face image from text, and close to none on 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 descriptions and face sketch pairs [4].

II. Materials and Method

We have found different papers regarding text-to-image conversion methods. One of them is Generating Images from Captions [5], 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 the Microsoft COCO dataset to accomplish these tasks. Figure 1 is an example of their work.

Fig. 1 Generating Images from Captions

Another paper, Generative Adversarial Text to Image Synthesis [6], uses the Caltech-UCSD birds database and the 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: learning a text feature representation that captures the essential visual details and using these features to synthesize a compelling image that a human might mistake for real. Examples of their work are shown in Figure 2.

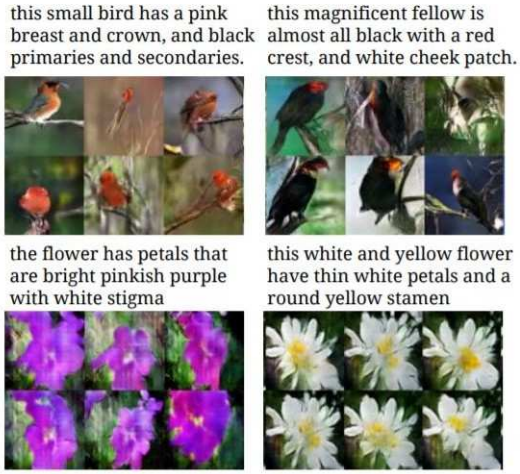


Fig. 2 Generative Adversarial Text to Image Synthesis

An attempt to generate face images from text descriptions can be seen in "Text2FaceGAN: Face Generation from Fine Grained Textual Descriptions" [7]. This paper leverages the CelebA dataset and generates a text-to-face dataset by auto-generating captions given the attributes of each image. These captions include descriptions of six sentences based on the structure of the face, facial hair, hairstyle, fine-grained face details, and accessories worn [7]. 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 is known to obtain outstanding results for the image retrieval task [7]. A representative output presented in the paper can be seen in Figure 3. Although the paper produces faces that align with the attributes in the descriptions, it generates low-resolution face images, making them difficult 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 the quality of image synthesis from text descriptions, a state-of-the-art approach was introduced by [8]. The model includes a layered attentional GAN, which can automatically select the condition at the word level for generating different parts of the image AttnGAN achieves a high-resolution image by combining the regional image vectors, generating new image features for each sub-region. 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 at the University of Queensland [9]. This paper utilizes the CelebA dataset and StyleGAN to achieve high-resolution 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 utilizes a multi-label classifier (T) to generate 40 facial attributes from free-form natural language descriptions.

Additionally, it employs an image encoder (E) from MobileNet to obtain image embeddings. Furthermore, it uses a pre-trained Style-GAN2 model to generate the final set of images. The paper utilizes a noise vector as an input to ensure diversity. Additionally, the paper proposes four noise vector manipulation techniques (differentiation, nonlinear re-weighting, normalization, and feature lock) to improve the performance of the model [9]. The literature introduced above provides 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. To perform Sketch generation from text, it is essential to sample from the distribution of our sketch images based on 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"[8]. AttnGAN consists of two essential 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 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 enable the generative networks to extract particular parts of the image conditioned on relevant words.

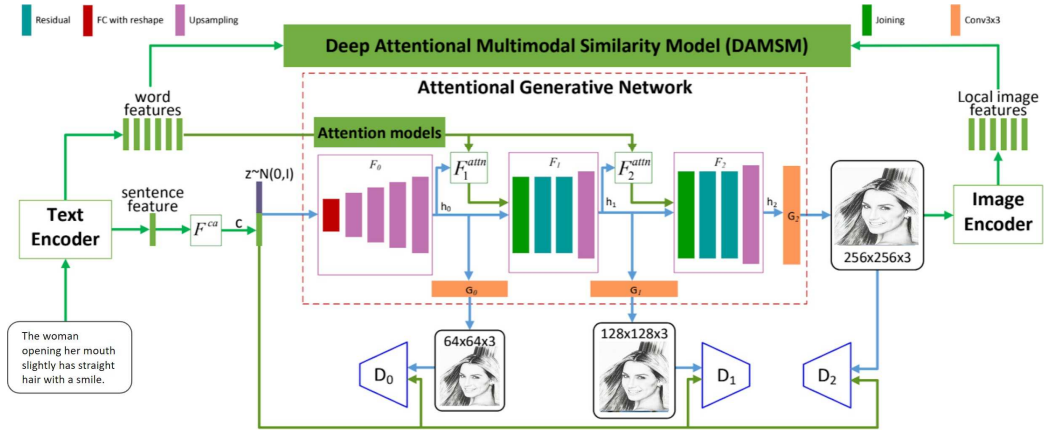


Fig. 5 Adapted AttentionGAN architecture

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 detail in Figure 6, utilizes features from the Text Encoder and the Image Encoder, and 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.

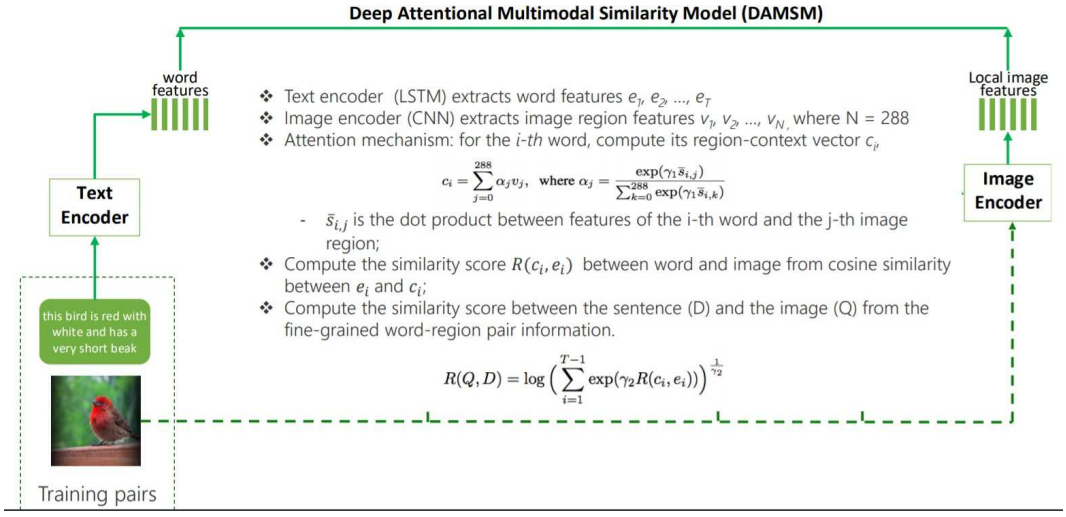


Fig. 6 Deep Attentional Multimodal Similarity Model Architecture

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 achieve a reasonable result, training for at least 30 epochs is required. Given this time constraint, we were limited in our experiments even with multiple AWS instances. Based on our experiments, we have found that with LSTM GRU type, 512 embedding dimensions, 0.25 gradient clipping, and 0.5 dropout works best for DAMSM. For the 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.

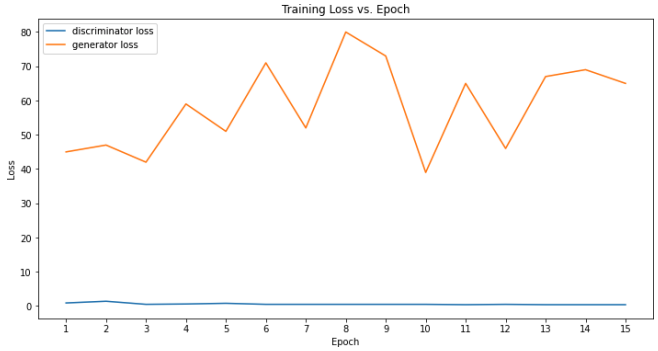


Fig. 7 Final Model Training

D. Dataset

Currently, there is no publicly available dataset of text and face sketches. Thus, we have created our own text and face sketch pair dataset based on 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 created by randomly selecting binary features from the CelebA dataset and incorporating 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. Sketch Generation

To generate sketch images from the RGB images in the CelebA dataset, we have trained a CycleGAN using the CUFS and CUFSF datasets. The CUFS dataset (consisting of the CUHK student dataset [4], the AR dataset [10], and the XM2VTS dataset [11] 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.

G. Baseline

Since there is no pre-existing dataset to train any model, generating the dataset itself serves as the baseline. The text is auto-generated from the 40 binary features in the CelebA dataset, as shown in Figure 8. For a generated text, we generate a sketch from the corresponding RGB CelebA image using the CycleGAN. To train CycleGAN, we utilized the images from the Celebrity dataset along with the sketches from the CUFSF dataset, as depicted in Figures 9 and 10.

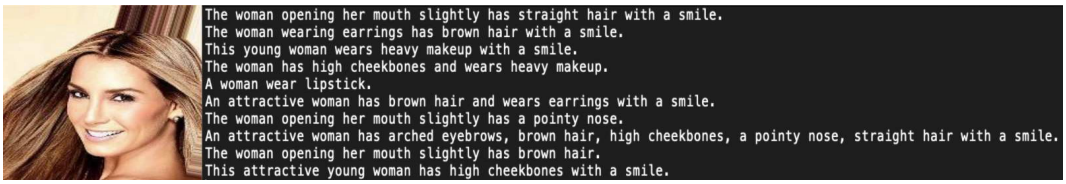


Fig. 8 Auto-generated Text Descriptions from CelebA Dataset

Fig. 9 Examples of CUFS dataset

Fig. 10 Examples of the CUFSF dataset

As this is an unpaired 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 to produce a decent result. An example output of the CycleGAN during its training procedure is depicted in Figure 11. As we mentioned previously, the caveat for GANs is that it is difficult to obtain good results from them, 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

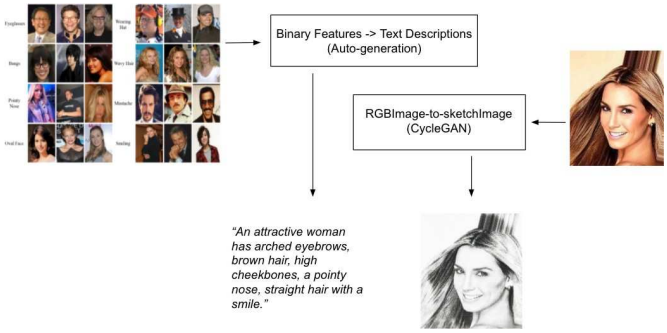


Fig. 12 Baseline/Dataset Generation

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 Divergence 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 [12]. This score focuses on the diversity of the images rather than the real image. The score is given as the following equation [12].

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

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 distribution $p(y) = E_{x\sim p_g}[p(y|x)]$.]. y is the label according to the inception network [12].

J. Fréchet Inception Distance (FID)

Fréchet Inception distance (FID) compares Inception embeddings (responses of the penultimate layer of the Inception network) of the real and generated images [12]. The distance (FID) is defined as follows.

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

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

In general, the Inception score estimates the quality and diversity of the collection of the generated images through the Inception V3 model. For the application to this task, the Inception score would be a good metric to evaluate the quality of synthesized images. Still, it could be harsh on grading the diversity of the photos since we only generate faces. FID score also uses the inception V3 model. Still, it extracts the CV-specific features of the input image and evaluates the closeness of the generated data distribution and the real data distribution. FID can be a good metric to assess the performance of the model for this task. Still, it may lack the ability to evaluate better models that generate a diverse set of 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 [13]-[16].

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 can capture facial characteristics such as "wavy hair", "high cheekbones", "opening mouth", "woman", and "smile".

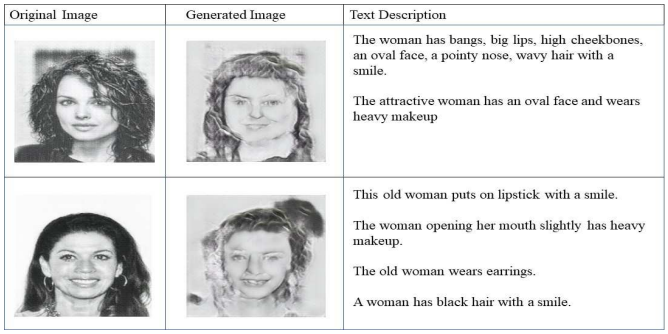


Fig. 13 Sample result of sketch generation from text description from test dataset.

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 the 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 was 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 model, 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 an attention map to observe which words are responsible for generating each section of the image. The corresponding region responsible for each text is marked in 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 the face structure of men and women.

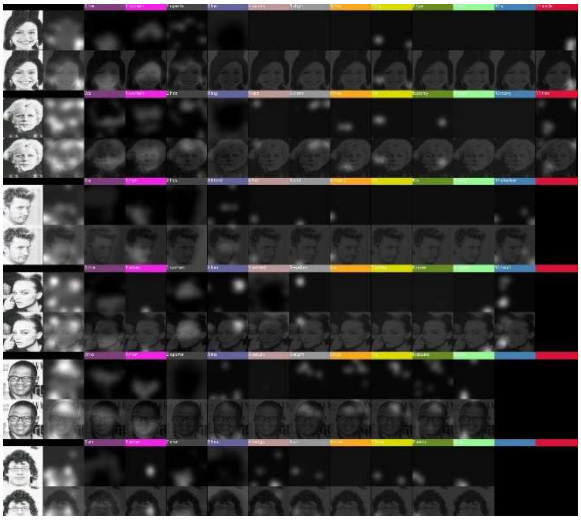


Fig. 14 Attention Map of DAMSM after 100 epochs

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.

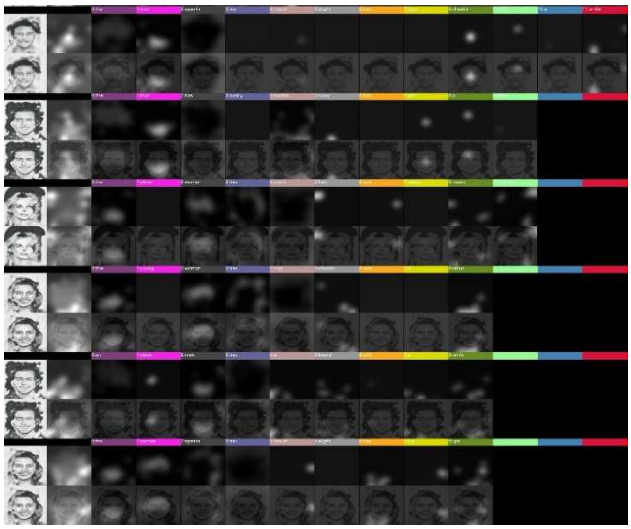


Fig. 15 Attention Map of AttnGAN after 10 epochs

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 the application of AttnGAN to generate sketch images successfully captures 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 results in low diversity and unrealistic images, having a significant impact on the model's 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] X. Chen, L. Qing, X. He, X. Luo, and Y. Xu, "FTGAN: A fully-trained generative adversarial networks for text to face generation," *arXiv preprint arXiv:1904.05729*, 2019.

[2] W. Zhang, X. Wang, and X. Tang, "Coupled information-theoretic encoding for face photo-sketch recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2011, doi:10.1109/CVPR.2011.5995324.

[3] Y. Wang *et al.*, "Text2Sketch: Learning face sketch from facial attribute text," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, doi: 10.1109/ICIP.2018.8451236.

[4] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, 2009, doi:10.1109/TPAMI.2008.222.

[5] E. Mansimov, E. Parisotto, J. L. Ba, and R. Salakhutdinov, "Generating images from captions with attention," *arXiv preprint arXiv:1511.02793*, 2016.

[6] 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.

[7] O. R. Nasir, S. K. Jha, and M. S. Grover, "Text2FaceGAN: Face generation from fine-grained textual descriptions," in *Proc. IEEE Int. Conf. Multimedia Big Data (BigMM)*, 2019, doi:10.1109/BigMM.2019.00-42.

[8] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks," *arXiv preprint arXiv:1711.10485*, 2017.

[9] T. Wang, T. Zhang, and B. C. Lovell, "Faces la Carte: Text-to-face generation via attribute disentanglement," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2021, doi:10.1109/wacv48630.2021.00342.

[10] A. M. Martinez and R. Benavente, "The AR face database," *CVC Tech. Rep. 24*, Jun. 1998.

[11] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," in *Proc. Int. Conf. Audio- and Video-Based Person Authentication*, 1999, pp. 72–77.

[12] K. Shmelkov, C. Schmid, and K. Alahari, "How good is my GAN?," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, doi: 10.1007/978-3-030-01216-8_14.

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

[14] H. Zhang and D. Metaxas, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, doi:10.1109/ICCV.2017.629.

[15] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, doi:10.1109/ICCV.2015.425.

[16] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, doi:10.1109/ICCV.2017.244.