Development and deployment of a generative model-based framework for text to photorealistic image generation

Sharad Pande [a], Srishti Chouhan [a], Ritesh Sonavane [a], Rahee Walambe [a,b], George Ghinea [c], Ketan Kotecha [a,b,*]

[a] Symbiosis Institute of Technology (SIT), Symbiosis International (Deemed University), India
[b] Symbiosis Centre for Applied Artificial Intelligence, Symbiosis Institute of Technology (SIT), Symbiosis International (Deemed University), India
[c] Department of Computer Science, Brunel University, London, UK

ARTICLE INFO

Keywords:
Text-to-image
Text-to-face
Face synthesis
GAN
AttnGAN

ABSTRACT

The task of generating photorealistic images from their textual descriptions is quite challenging. Most existing tasks in this domain are focused on the generation of images such as flowers or birds from their textual description, especially for validating the generative models based on Generative Adversarial Network (GAN) variants and for recreational purposes. However, such work is limited in the domain of photorealistic face image generation and the results obtained have not been satisfactory. This is partly due to the absence of concrete data in this domain and a large number of highly specific features/attributes involved in face generation compared to birds or flowers. In this paper, we propose an Attention Generative Adversarial Network (AttnGAN) for a fine-grained text-to-face generation that enables attention-driven multi-stage refinement by employing Deep Attentional Multimodal Similarity Model (DAMSM). Through extensive experimentation on the CelebA dataset, we evaluated our approach using the Frechet Inception Distance (FID) score. The output files for the Face2Text Dataset are also compare with that of the T2F Github project. According to the visual comparison, AttnGAN generated higher-quality images than T2F. Additionally, we compare our methodology with existing approaches with a specific focus on CelebA dataset and demonstrate that our approach generates a better FID score facilitating more realistic image generation. The application of such an approach can be found in criminal identification, where faces are generated from the textual description from an eyewitness. Such a method can bring consistency and eliminate the individual biases of an artist drawing the faces from the description given by the eyewitness. Finally, we discuss the deployment of the models on a Raspberry Pi to test how effective the models would be on a standalone device to facilitate portability and timely task completion. ©

1. Introduction and scope

Reed et al. [1] first introduced text-to-image synthesis in 2016, and it is a fundamental and novel research area in computer vision [2]. It is similar to reverse image captioning in that it aims to create natural images from input sentences. Text-to-image synthesis, including image caption, explores the visual semantic process of the human brain by mining the connection between text and image. Furthermore, it has enormous potential for art production, computer-aided design [3], image searching, and other areas such as image analysis of gold immunochromatographic strip [4–6]. Recently, methods for text-to-image synthesis based on Generative Adversarial Networks (GANs) [7] have been suggested. It is common to encode the entire text meaning into a global sentence vector as a prerequisite for GAN-based image production [1,8–11]. Despite its promising results, conditioning a GAN only on the global sentence vector has certain limitations and is inadequate in taking into consideration the crucial fine-grained details at the word stage. This limitation is even more evident when creating complex scenes like those in the COCO dataset [12] or the CelebA dataset [13].

Text-to-face synthesis is a subdivision of text-to-image synthesis that aims to create face images from human descriptions. Text-to-face synthesis, similar to text-to-image synthesis, has two key goals: (1) to produce high-quality images and (2) to generate images that correspond to the input descriptions.

* Corresponding author at: Symbiosis Institute of Technology (SIT), Symbiosis International (Deemed University), India.
E-mail address: head@scaai.siu.edu.in (K. Kotecha).

1.1. Related work

Text-to-face synthesis is divided into two categories: (1) text-to-image synthesis and (2) text-to-face synthesis. Fig. 1 summarises the previous work carried out in both these categories.

- Text-to-image:

Despite the fact that there are a variety of networks for text-to-image synthesis tasks, the majority of them are built on the encoder-decoder structure and conditional GAN [14]. Text-encoder and image-decoder are used in this encoder-decoder system. The text encoder converts input definitions into semantic vectors, which are then decoded into natural images by the image decoder. Text-to-image synthesis has two key goals: to create high-quality images and to create images that complement the provided descriptions. These two goals serve as the foundation for all advances in text-to-image synthesis.

Text-to-image synthesis research in its early stages has primarily focused on improving the quality of produced photographs. Reed et al. proposed the text-to-image challenge for the first time in 2016 and created two end-to-end networks focused on conditional GAN to complete it [1]. Reed used a pre-trained Char-CNN-RNN network for text encoding and a DCGAN-like network [15] for image decoding to create natural images from vectors. Following that, several scholars made advancements as a result of his study [16]. Zhang et al. published one of the most influential studies on this subject, proposing a two-stage network called StackGAN [8] to solve the problem and for producing high-quality photos and improving the Inception Score. Later studies [3,10,17,18] inherit this network as well.

Researchers gradually concentrated on achieving another goal: enhancing the resemblance between input text and produced photos because the network had already shown the ability to produce realistic images. Reed et al. suggested a network for generating images from a box that was created first. This approach assisted in producing more precise data on the output photographs [9]. A GAN network built on a related concept was also created by [19]. Sharma et al., on the other hand, used dialogue to aid interpretation of the description, which allows synthesising visuals that are more relevant to the input document [20].

Dong et al. suggested a method for creating new images based on the input picture and explanations, which would produce images that complement the input descriptions [16]. They also proposed Image-Text-Image (I2T2I), a new training approach that combines text-to-image and image-to-text (image captioning) synthesis to increase text-to-image synthesis accuracy [21]. Attention processes have already made significant progress in the text- and image-related activities [22–25] and they are now being used in GANs to generate text-to-image conversions. [3] constructed Attn-

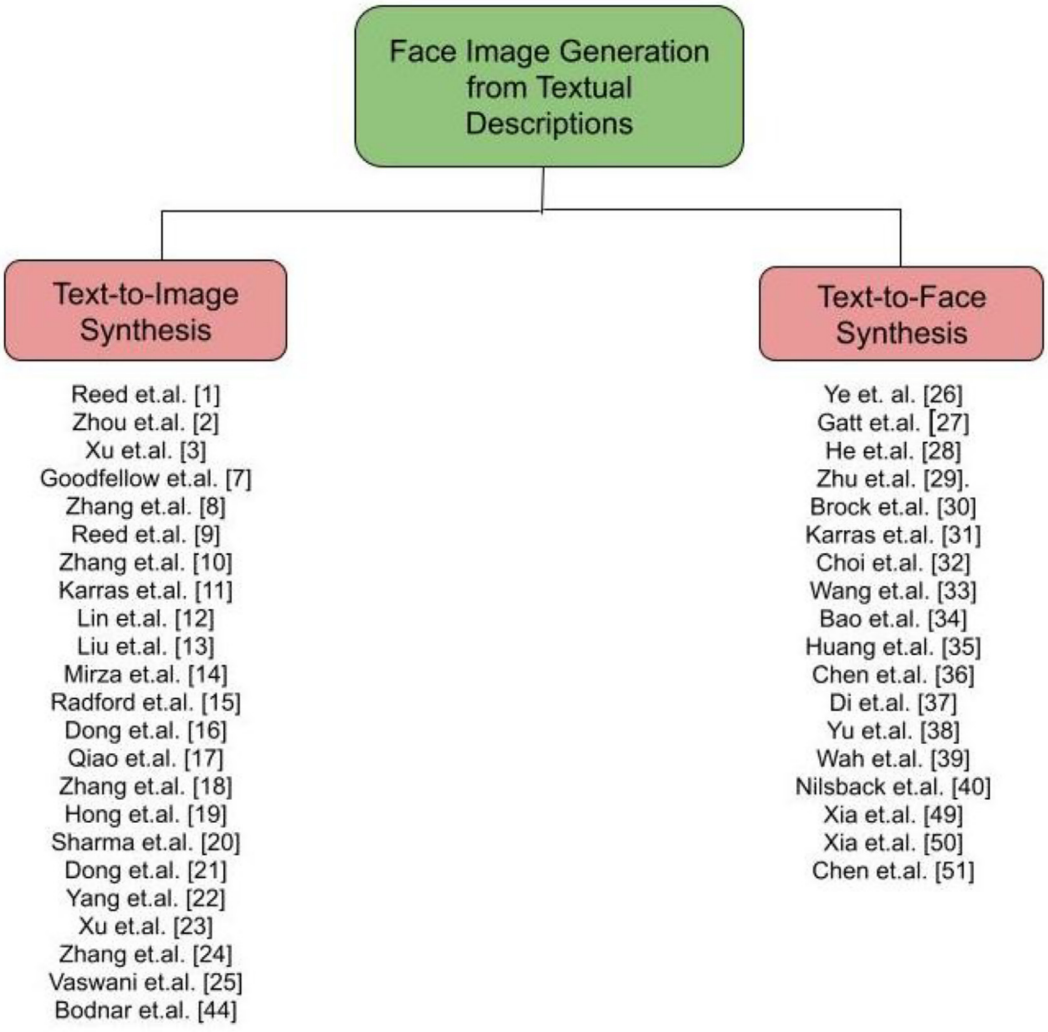

Fig. 1. Previous work in Text-to-Face synthesis.

GAN to create an attention mechanism that allows GANs to create fine-grained high-resolution photographs from natural language descriptions. MirrorGAN [17] is a text-to-image-to-text network suggested by Qiao et al., which uses a global–local collaborative focus paradigm. [18] proposed a visual-semantic similarity measure as an aid to measurement metrics since there are no available criteria on how the produced images represent the input descriptions. These findings indicate a pattern in which researchers are increasingly concentrating on improving the accuracy of produced images and input sentences. With this, we can use it for scripts-to-storyboard, text-to-architecture and much more.

- Face Synthesis:

Image synthesis has been a popular subject in deep learning since Goodfellow suggested GAN in 2014 [7,26]. Face synthesis is a common research area since there are two broad-scale public datasets: CelebA [13] and Face2Text [27]. Almost all state-of-the-art networks, like networks based on GAN and networks based on both conditional GAN and conditional GAN, demonstrates their model's dominance on face synthesis (such as PGGAN [11], DCGAN [15,28] CycleGAN [29], BigGAN [30], StyleGAN [31], StarGAN [32] to name the few). With the advancement of these networks, the quality of produced face images is steadily improving. Some networks now generate 1024×1024 face pictures, which is far greater than the face dataset's initial picture resolution. These models attempt to learn a mapping between a noise vector and real face images that follows the Normal distribution. However, they are unable to command the network to produce the particular face picture that they need.

Face synthesis has derived many interesting applications about faces using conditional GAN, such as translating edges to natural face images [33], exchanging the attributes of two face images [34], generating a positive face from the side face [35], generating a full face from the eyes' region only [36], from face attributes to sketches to natural face images synthesis [37], and face in space [38]. By applying a condition vector to the synthesised face images, such networks attempt to monitor them and produce face images that meet the needs of various circumstances. The activities that use the input descriptions as the control condition are identical to text to face synthesis.

In the context of face generation from the textual description, one of the most relevant applications is to develop criminal images from the textual description from an eyewitness. In the public safety context, this task has more relative values than text-to-image synthesis. Drawing an image for a criminal solely relying on eyewitness descriptions is a daunting process that takes technical knowledge and extensive practice. Additionally, the individual biases of the eyewitness as well as the artist may creep into the process. E.g., different artists may have a different notion of 'attractive' or 'dark skin' based on their social and ethnic background. Such biases may bring inconsistency in the images created and delay the process of finding the criminals. On the other hand, a person who is not an artist can also easily produce photorealistic faces of criminals based on eyewitness reports using a well-trained text-to-face model.

To address this problem, we employ the Attentional Generative Adversarial Network (AttnGAN) [3], which enables fine-grained text-to-image creation through attention-driven multi-stage refinement. The model is made up of two unique elements. The attentional generative network is the first part, in which the Generator generates an attention system that enables it to draw different sub-regions of the image by concentrating on words that are more relevant to the sub-region that is being drawn. In addition to the natural language summary being encoded into a global sentence vector, each word in the sentence is encoded into a word vector. The other part of the AttnGAN includes a Deep Attentional Multimodal Similarity Model (DAMSM). With the aid of an attention mechanism, the DAMSM computes the similarity between the generated image and the sentence using both global sentence-level details and fine-grained word-level information. Consequently, the DAMSM modifies the Generator's preparation by adding a finer-grained image-text matching loss. We consider birds [39], flowers [40], CelebA [13] and Face2Text [27] datasets for the study and experimentation.

1.2. Contribution and novelty

In summary, the main contributions of this work are:

(i) The study, comparison and analysis of various GAN models for photorealistic image generation from textual descriptions by experimentation on birds, flowers, and human faces datasets.
(ii) Implementation of the Text-To-Face synthesis using AttnGAN through attention-driven multi-stage refinement for photorealistic face image generation and optimisation of the model by employing DAMSM loss. A model architecture is proposed based on the AttnGAN employing the DAMSM loss.
(iii) Implementation of the trained models on a standalone Raspberry Pi device to ensure more portability, useability and accessibility of such an approach.

1.3. Novelty

1. Experimentation for identifying various aspects of Generative models for photorealistic face image generation. This includes the in-depth analysis of the effect of FID scores and DAMSM loss on image quality and realism.
2. Comparison and analysis of the results obtained with existing methods on CelebA dataset.
3. Generation of distinct variations in the images as a result of semantic alterations in the input text.
4. Implementation on a standalone portable hardware system for easy application and usability.

1.4. Outline

The paper is organised as: Section 2 presents the methods and the background of the architectures employed in detail, followed by Section 3, which describes the experiments, evaluations, and results. Section 4 presents the discussion on experiments and the results. The paper concludes with an overview and future scope in Section 5.

2. Methods and background

In this section, we present AttnGAN and the further application of that for text to face synthesis. We begin by explaining how Generative Adversarial Networks (GAN) function. Then we describe AttnGAN and its DAMSM network to carry out text encoding and compute the attention map. This attention map is then utilised for the task of generating images from their textual description. Further, we describe how AttnGAN is helpful for our problem statement. Finally, we describe the FID Score and evaluate our model.

2.1. Generative Adversarial Networks (GAN)

GAN stands for Generative Adversarial Networks, and it is a framework for learning a function or program that can produce

samples that are quite similar to samples taken from a specified training distribution. GANs have become popular very recently. The general architecture of a GAN [7] consists of a Generator(G) and a Discriminator(D). Both the Generator and Discriminator are separate neural networks. A random input(noise) is given to the Generator and it tries to produce an image close to the actual image. The output of the Generator is then given to the Discriminator. The Discriminator tries to tell the difference between natural and synthetic training data, whereas the Generator tries to deceive the Discriminator. The Discriminator updates the weights depending on whether it predicts the image generated by the Generator as real or fake. If it predicts the image to be fake, then an update in the weights takes place. The duty of the Generator here is to keep on producing images that seemingly look real. It does so till the time that the Discriminator predicts them as real images. So essentially, the Generator and Discriminator participate in a minimax game. Equations (1) and (2) from [41] describe this minimax game.

$$J^{(D)} = -\frac{1}{2}E_{x\sim P_{data}} logD(x) - \frac{1}{2}E_z log(1 - D(G(z))) \quad (1)$$

$$J^{(G)} = -J^{(D)} \quad (2)$$

P_{data} is the probability distribution of given data, and J (D) is the discriminator cost, and J (G) is the generator cost. The Nash Equilibrium of this game, according to Goodfellow et al. [7], is when samples generated by G are indistinguishable from samples derived from training data (provided G and D have sufficient capacity).

2.2. StackGAN

This GAN [8] is typically used for synthesising images from textual description. It breaks down the text-to-image generation process into two stages, as mentioned below:

1) Stage-I GAN

Stage-I GAN focuses on drawing only rough shapes and appropriate colors from the textual definition. It creates a low-resolution image by drawing the context layout from a random noise vector. It generally produces 64x64 images.

2) Stage-II GAN

Stage-II GAN is built upon Stage-I GAN results, and so it produces high-resolution images. Low-resolution images generated by Stage-I GAN are generally devoid of realistic object parts and might have distortions of shape. The Stage-II GAN takes into consideration the text ignored in Stage-I to generate images with more natural details. It generates 256x256 images.

2.3. PGGAN

PGGAN [11] is short for Progressively Growing GAN. PGGAN is used to produce ultra-high-resolution images by increasing the network layers as training goes on by first training a model to generate 4x4 image and add layers to generate 8x8, 16x16 images and so on.

The most significant difference between PGGAN and StackGAN is that the network structure of the latter is fixed. However, in PGGANs, as the training progresses, the network structure continues to change. The most significant benefit of doing this is that most iterations are done at lower resolutions, and the training speed is faster than traditional GANs.

2.4. AttnGAN

AttnGAN or Attentional Generative Adversarial Network [3] is an attention-driven architecture that enables text-to-image conversion. The architecture involves multiple stages for the generation of fine-grained images. It generates high-quality images by dividing an image into various subregions and then focusing on specific words from the caption relevant to a particular subregion of the image.

The models that have so far been used for text-to-image conversion use the entire description and convert it to a vector which is then used for image generation. In this model, instead of the whole sentence, we focus on its constituent words to generate various image subregions. This ensures a generated image that is visually closer to the actual image. Different words are used to produce different parts of the final image according to the sub-region that they are most relevant to. The detailed architecture for AttnGAN is shown in Fig. 2.

The text description, containing T words, is input to the text encoder. The text encoder is a bidirectional LSTM. This means that the input caption is trained on two LSTMs instead of the usual one LSTM. Essentially this does the job of concatenating the hidden states from the forward and backward direction for all timesteps and outputs a final hidden state. This final hidden state in the case of this architecture is the sentence feature, represented by $\overline{e}(\overline{e} \in \mathbb{R}^D)$. Here$D$ represents the working dimension for the words. Since there are T words, hence another matrix $e$with the dimension $e \in \mathbb{R}^{D\times T}$represents the word features. In a nutshell, sentence features may be considered as the final hidden state, while the word features are the hidden states from all timesteps.

The sentence features are passed on to F^{ca} for conditional augmentation. F^{ca} is modelled as a neural network. All the equations in this section are based on the mathematical discussion presented in [3].

The output after Conditional Augmentation, c, is given by:

$$c = \mu + \sigma\varepsilon, \varepsilon \sim \mathcal{N}(0, I) \quad (3)$$

Since the same description can describe several images, $\varepsilon_{(noise)}$is added here to introduce variation in the generated images.

Typically, the input to the Generator in a GAN is only a noise vector (z). But since to generate the final image, the Generator needs to be conditioned on the input description, so here we use Conditional GANs. Accordingly, $c$and z(noise vector) are concatenated and are fed as input to the generator network.

The architecture can be considered to have $m$generators.

F_0 is responsible for most of the upsampling. The scale factor for upsampling is 2. F_0 does not use word-level features. The context vector at this stage, h_0, is given by:

$$h_0 = F_0(z, F^{ca}(\overline{e})) \quad (4)$$

The output from $F_0(i.e.h_0)$ along with the word features (e) is taken as input by the attention network. To do this, a perceptron layer is added to transform the word features into a common semantic space of image features. This may be represented as: $e' = Ue$, where $U \in \mathbb{R}^{\widehat{D}\times D}$. Here $\widehat{D}$ represents the network's internal working dimension.

Together with h_0, e' is given as input to the attention network. Thus, we get a word-context vector for every subregion.

The word-context vector may be understood as a score to relate all T words to all N subregions and is a measure of how relevant a word would be to a particular region. This is how specific words are selected for generating specific regions of the final image. The word context vector for the j^{th} subregion is given as:

$$c_j = \sum_{i=0}^{T-1} \beta_{j,i} e'_i, \text{ where } \beta_{j,i} = \frac{\exp\left(s'_{j,i}\right)}{\sum_{k=0}^{T-1} \exp\left(s'_{j,k}\right)} \text{ and } s'_{j,i} = h_j^T e'_i \quad (5)$$

On doing this for each region, we get the output for the attention network, which is:

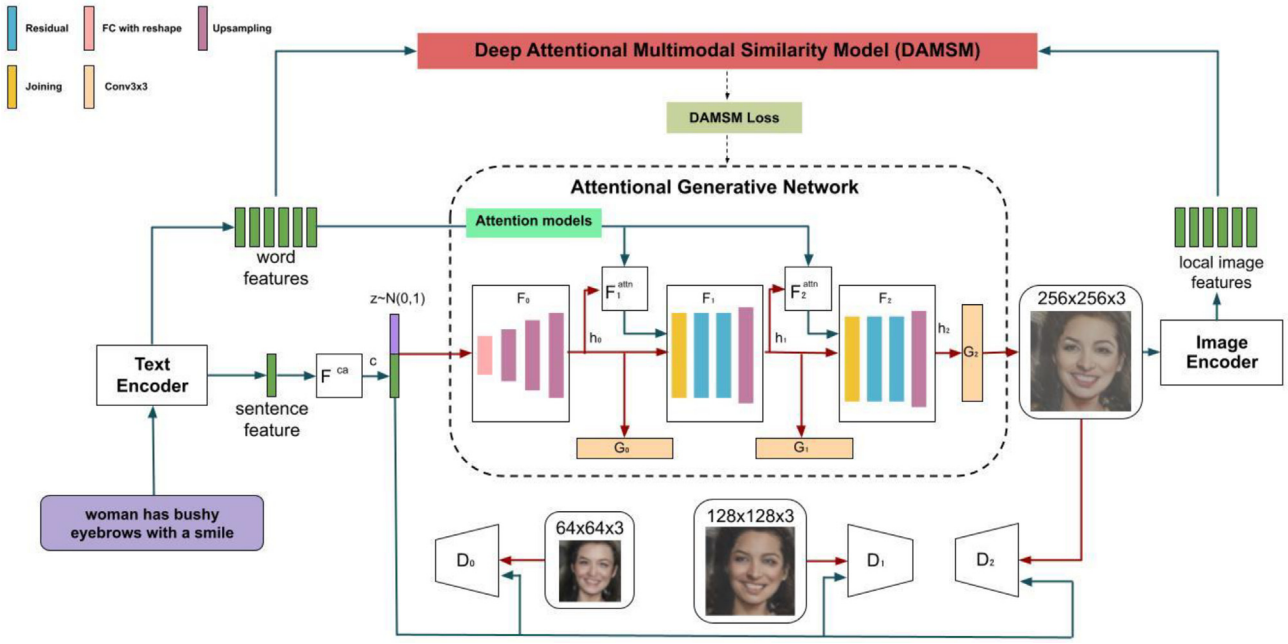

Fig. 2. The Architecture for AttnGAN for text-to-face synthesis. Each attention model automatically retrieves the conditions (i.e., the most critical word vectors) for generating various sub-regions of the image; the DAMSM provides the fine-grained image-text matching failure for the generative network.

$$F^{attn}(e,h) = (c_0, c_1, ..., c_{N-1}) \in \mathbb{R}^{\widehat{D}\times N} \tag{6}$$

For F_1, there are two inputs, h_0 from F_0 and the word-context vector from the attention network. It consists of residual blocks, which make the network deeper, and an upsampling layer. It uses word-level features from F_1^{attn}. Here F_i^{attn} is the attention model at i^{th} stage of AttnGAN. The context vectors henceforth may be generalised as:

$$h_i = F_i(h_{i-1}, F_i^{attn}(e, h_{i-1})) \text{for} i = 1, 2, ..., m-1; \tag{7}$$

The second attention network, F_2^{attn} and F_2 have the exact functionality and structure as F_1^{attn} and F_1. They only differ in inputs.

The number of generators can be adjusted according to the size of the image to be produced. The more the number of generators, the more is the size of the image.

Each of the F generators is associated with a G generator. They consist of convolutional blocks which bring down the number of channels to, i.e., an RGB image. The generated image, $\widehat{x_i}$, is given by:

$$\widehat{x_i} = G_i(h_i) \tag{8}$$

The overall generator loss is the summation of all the$m$generators present in the network this is given by

$$\mathcal{L}_G = \sum_{i=0}^{m-1} \mathcal{L}_{G_i} \tag{9}$$

The discriminators placed after the generated image represent whether or not the generated image justifies the input caption. For this, one of its inputs is c, which is the output after conditional augmentation. The other input is the generated image.

Losses:

The adversarial loss for i^{th} generator, G_i, is:

$$\mathcal{L}_{G_i} = -\frac{1}{2}\mathbb{E}_{\widehat{x_i}\sim \mathcal{P}G_i}\Big[log(D_i(\widehat{x_i})] - \frac{1}{2}\mathbb{E}_{\widehat{x_i}\sim \mathcal{P}G_i}[log(D_i(\widehat{x_i}, \overline{e})] \tag{10}$$

Here the unconditional loss tells whether or not the image is real, and the conditional part tells whether the generated image and the input description belong to the same pair.

Cross entropy loss for Discriminator,

$$\mathcal{L}_{D_i} = -\frac{1}{2}\mathbb{E}_{x_i\sim \mathcal{P}data_i}\Big[log(D_i(x_i)] - \frac{1}{2}\mathbb{E}_{\widehat{x_i}\sim \mathcal{P}G_i}\Big[log(1 - D_i(\widehat{x_i})\big] + -\frac{1}{2}\mathbb{E}_{x_i\sim \mathcal{P}data_i}\Big[log(D_i(x_i, \overline{e})] - \frac{1}{2}\mathbb{E}_{\widehat{x_i}\sim \mathcal{P}G_i}\big[log(1 - D_i(\widehat{x_i}, \overline{e})\big] \tag{11}$$

The attentional network's final objective function is provided by:

$$\mathcal{L} = \mathcal{L}_G + \lambda\mathcal{L}_{DAMSM}, \text{ where } \mathcal{L}_G = \sum_{i=0}^{m-1}\mathcal{L}_{G_i} \tag{12}$$

2.5. Deep attentional multimodal similarity model (DAMSM) model

The paper [3] also introduces a deep attentional multimodal similarity model to determine a loss indicating how the generated fine-grained image matches its corresponding text/caption. This is used for training the Generator.

DAMSM tries to check whether the generated images follow the written textual description. It does so with the help of two neural networks. Essentially, it tries to calculate the similarity between the image and the text to generate an image with better detailing.

An image encoder is required to compute the DAMSM loss. It takes the generated image as input. It is based on the Inception-v3 model [42], which has been pre-trained on ImageNet [43]. Then, from the "mixed_6e" layer of Inception-v3, we extract the local feature matrix $f \in \mathbb{R}^{768\times 289}$ (reshaped from $768 \times 17 \times 17$). The feature vector of a local image area is represented by each column of f. The local function vector has a dimension of 768, and the picture has289regions. Meanwhile, the global function vector $\overline{f} \in \mathbb{R}^{2048}$ is derived from Inception-v3's last average pooling layer. All equations in this section are proposed in [3].

Finally, as seen in below Eq, we bring the image features to a common space.

$$\upsilon = Wf, \upsilon \in \mathbb{R}^{D\times 289} \tag{13}$$

$$\overline{\upsilon} = \overline{Wf}, \overline{\upsilon} = \mathbb{R}^{D} \tag{14}$$

After this, the similarity matrix is computed, which is given by:

$$s = e^T v, s \in \mathbb{R}^{T \times 289} \tag{15}$$

It was found that normalising the similarity matrix gives better results, so the normalised matrix is:

$$\overline{s_{i,j}} = \frac{exp(s_{i,j})}{\sum_{k=0}^{T-1} exp(s_{k,j})} \tag{16}$$

Here, a region context vector,c_i, is calculated. This is different from the word-context vector. In the word-context vector, we were looking at all the words and estimating how relevant they would be to a particular region so that they may be used to generate that sub-region. As opposed to that, in the regional context vector (calculated for a word), we look at a single word at a time and look at all the sub-regions. This is done to understand whether that particular word was significant in the generation of a particular subregion. The region context vector is defined as follows:

$$c_i = \sum_{j=0}^{288} \alpha_j v_j, \quad where \quad \alpha_j = \frac{exp(\gamma_1 \overline{s_{i,j}})}{\sum_{k=0}^{288} exp(\gamma_1 \overline{s_{i,k}})} \gamma_1$$
$$: \ attention \ \ scaling \ \ factor \tag{17}$$

The relevance between i^{th}word and the image is also calculated. This tells us how important each word is in the generation of the image:

$$R(c_i, e_i) = \frac{c_i^T e_i}{||c_i|| ||e_i||} \tag{18}$$

Finally, the attention-driven image-text matching score between the image and its text description is given by:

$$R(Q, D) = log\left(\sum_{i=1}^{T-1} exp(\gamma_2 R(c_i, e_i))\right)^{\frac{1}{\gamma_2}}$$
$$\gamma_2 : word - to - region \ \ context \ \ pair \ \ vector \tag{19}$$

A similar score is calculated on the sentence level as well. It is given by :

$$R(Q, D) = \frac{\overline{v}^T \overline{e}}{(||\overline{v}|| ||\overline{e}||)} \tag{20}$$

The final DAMSM loss is given by:

$$\mathcal{L}_{DAMSM} = L_1^w + L_2^w + L_1^s + L_2^s \tag{21}$$

$$\mathcal{L}_1^w = -\sum_{i=1}^{M} logP(D_i|Q_i) \tag{22}$$

$$\mathcal{L}_2^w = -\sum_{i=1}^{M} logP(Q_i|D_i), \text{where} \ \ P(Q_i|D_i)$$
$$= \frac{\exp(\gamma_3 R(Q_i, D_i))}{\sum_{j=1}^{M} exp(\gamma_3 R(Q_j, D_i))} \tag{23}$$

here $P(D_i|Q_i)$ is the posterior probability of the sentence D_i being matched with the image Q_i. It is given by

$$P(D_i|Q_i) = \frac{\exp(\gamma_3 R(Q_i, D_i))}{\sum_{j=1}^{M} exp(\gamma_3 R(Q_i, D_j))} \tag{24}$$

3. Experiments

In this section, we begin by studying the important components of StackGAN and PGGAN to understand the architecture for text-to-image synthesis. We also look at AttnGAN's key components, such as the attentional generative network and the DAMSM. Additionally, we analyse why AttnGAN is a better architecture for complex scenes by implementing a specific application, namely, text-to-face synthesis and implementing it in Raspberry Pi.

3.1. Experiments on CUB-200 and Oxford-102 flowers dataset

To understand the architecture for generating images from their textual descriptions, we experimented with StackGAN and PGGAN [44]. The experiments were carried out using the Caltech CUB-200 dataset [39] and Oxford-102 Flower's dataset [40], for which the results are shown in Figs. 3, 4 and 5.

Although StackGAN and PGGAN work well on datasets of birds and flowers, they cease to produce visually satisfactory results on complex datasets like the COCO dataset or CelebA dataset. This can be observed from the results provided in previous studies [45].

3.2. Experiments on CelebA and Face2Text datasets

While StackGAN and PGGAN lack performance on the COCO dataset, AttnGAN has been used on the COCO dataset [12]. This dataset on [3] produced an Inception Score of 25.89 ± 0.47 as opposed to 8.45 ± 0.03 and 9.58 ± 0.21, respectively, on the former two architectures. Based on this, we deciphered that AttnGAN would work well for the Faces dataset as well.

Initially, our methodology is evaluated using CelebA datasets. We adopted and preprocessed the CelebA and Face2Text dataset

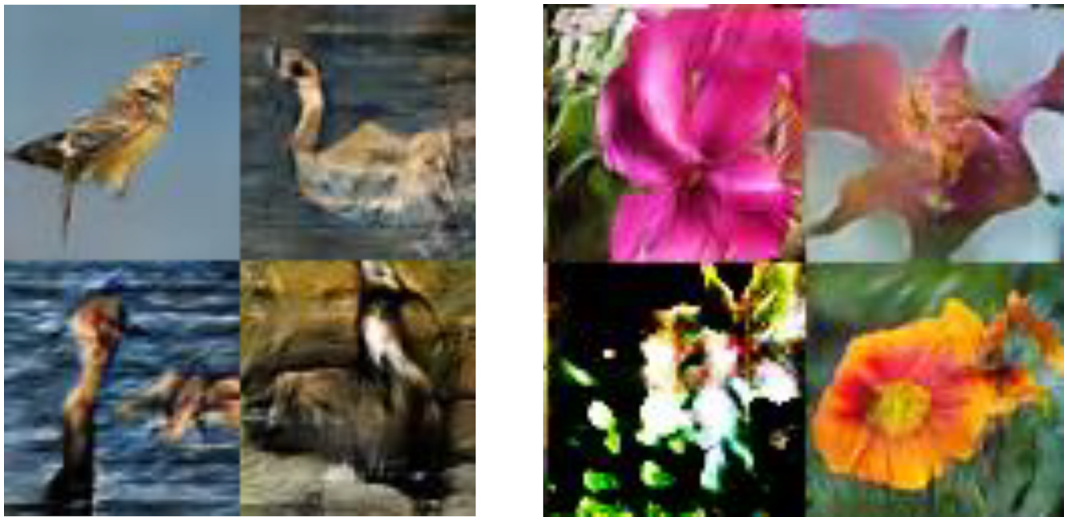

Fig. 3. Some examples of our generated images (64x64) by StackGAN Stage-I on Caltech CUB-200 and Oxford-102 datasets.

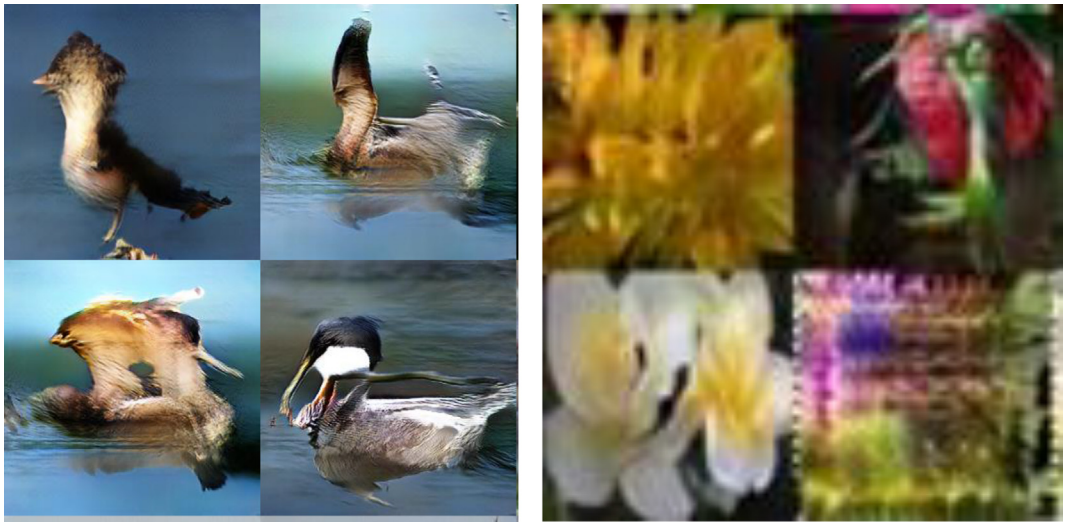

Fig. 4. Some examples of our generated images (256x256) by StackGAN Stage-II on Caltech CUB-200 and Oxford-102 flowers dataset.

Fig. 5. Some examples of our generated images by the PGGAN Oxford-102 flowers dataset.

using the methods described in [3] to assess the efficacy of our approach.

The Face2Text Dataset [27] was utilised as an experiment to see how AttnGAN could handle generating images for the face dataset. The dataset contains 400 images, the majority of which have three captions per image. We later reduced this to 2 captions per image by increasing the number of words in a sentence and thereby reducing the number of captions to be tested by the Generator during training. The output files were compared with that of the T2F Project on Github [45]. According to the visual comparison, AttnGAN generated higher-quality images than T2F (as shown in Fig. 6)

Since our results on the Face2Text dataset were visually promising, as seen from Fig. 6, we carried out further experiments with a bigger dataset, namely the CelebA dataset [13].

Fig. 6. The AttnGAN model generates the images on the left and the right images are generated examples from StackGAN and PGGAN models from [45], trained and tested it on the Face2Text dataset.

The Celeb Faces Attributes Dataset (CelebA) [13] is a large-scale face attribute dataset. It has over 200 K celebrity images and 40 attribute annotations for each image. The images in this collection cover a wide range of poses as well as the clutter in the background. CelebA has rich annotated information. We used 10.2 k images from the CelebA dataset to train the AttnGAN model. Since the dataset lacks official captions, captions were sourced from [46]. Each image has ten captions that cover all of the image's attributes.

An attention map for each word in the input statement, as shown in Figs. 7 and 8, is generated. In attention maps, words that are of use while producing a particular sub-region are highlighted. In the case of text-to-face, these include the words that describe the attributes of the face. When generating images, this shows where the network would concentrate with each word. When responding to certain terms, the induced attention maps essentially fit the concentrating region of the human brain. The generated face images have a high level of continuity with the input sentences. However, sometimes the attention maps fail to represent the captions accurately, as shown in Fig. 9.

The DAMSM Model was initially trained for each dataset until no significant changes in the sentence and word loss were observed for real image-text pairs. As a result, the image and text encoders learned how to extract global feature vectors from produced images and text descriptions. Further, as the attention GAN model was being trained, this pre-trained DAMSM model computed the $\mathcal{L}_{DAMSM}$ loss for each iteration.

The DAMSM loss is governed by a parameter λ, and the total loss of the model is given by the equation (12). To test $\mathcal{L}_{DAMSM}$ The value of λ is tuned from $\lambda = 0$ to 5. The results obtained for various λvalues are shown in Fig. 10.

These results show that appropriately raising $\mathcal{L}_{DAMSM}$ weight results in higher-quality images that are better trained on the given input descriptions. This is because increased $\mathcal{L}_{DAMSM}$ provides word-level matching information, which helps train the Generator in a better way. The CelebA dataset was also trained on$\lambda = 50$. But each time, the training resulted in a mode collapse which was unlike the case when AttnGAN was trained on the COCO dataset by [3].

This work aims not only to generate better quality (having more similarity to the textual description) images but also the images that retain realism and are visually more realistic. As can be seen from Fig. 10, the images generated for $\lambda = 3$are more realistic than $\lambda = 5$.

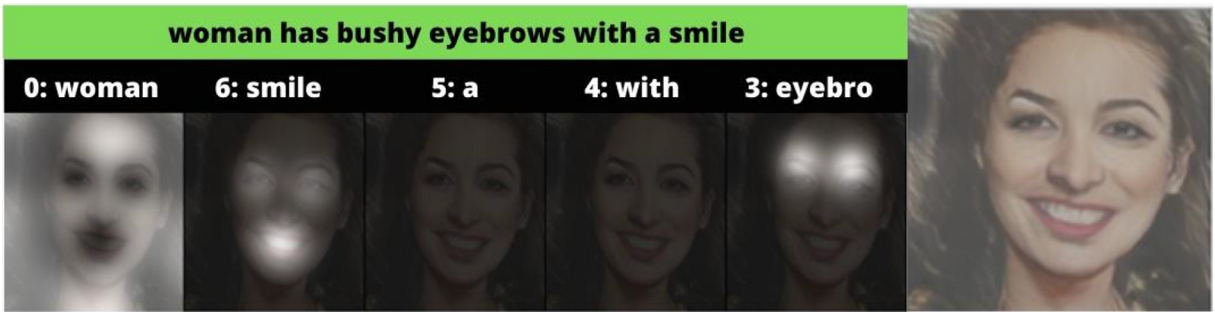

Fig. 7. Attention Maps of the Generated example of the text-to-face synthesis. The image shown has the input description as "woman has bushy eyebrows with a smile."

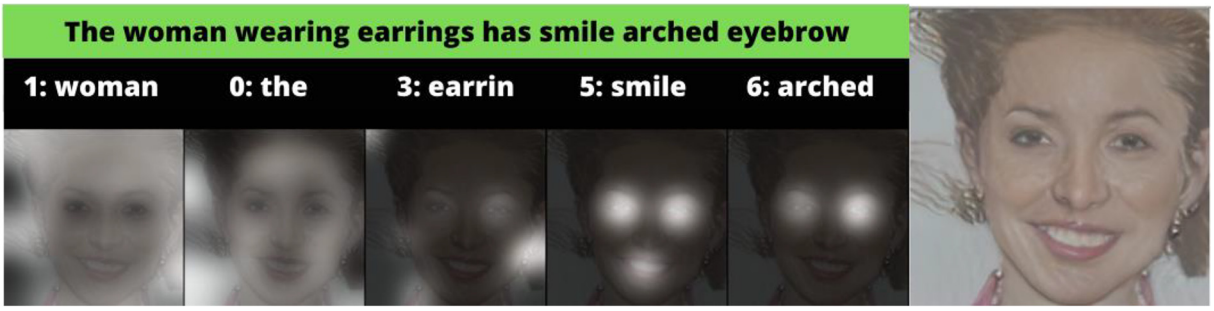

Fig. 8. Attention Maps of the Generated example of the text-to-face synthesis. The image shown has the input description as "The woman wearing earrings has smile arched eyebrow".

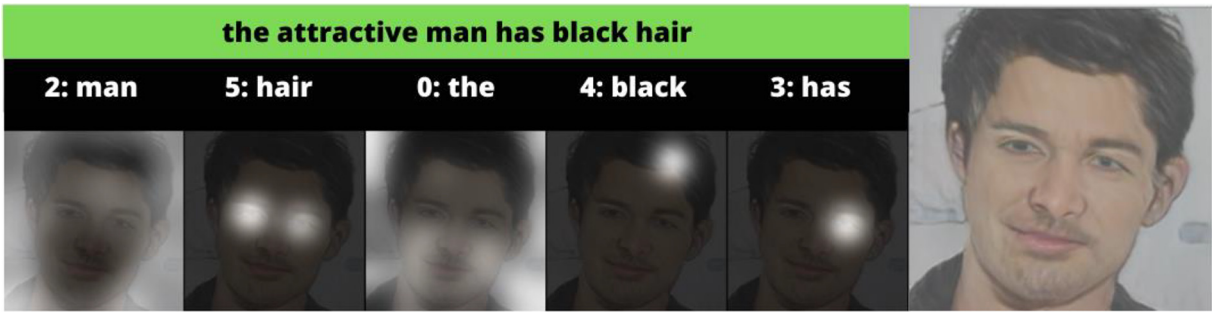

Fig. 9. The image shown has the input description as "the attractive man has black hair". It is observed in the image that the hair attribute has not been correctly represented through its attention map.

3.3. Evaluation

The FID score [47] is used to assess the image consistency of synthetically generated faces. Text-to-image synthesis, in general, uses Inception Score as a metric. Standard practice is to use a pre-trained Inception-V3 network that is fine-tuned on a specific images dataset to measure Inception Score to determine network outcomes. This is reported in [3] for the CUB dataset. However, there is no pre-trained Inception-V3 model for the face dataset. As a result, we switched to the FID score, which is another often-used metric for measuring image synthesis and can be thought of as a more powerful variant of the Inception Score (IS) as it is more robust to noise than IS.

FID is a metric for comparing the resemblance between two image datasets. It is found to associate well with human visual content judgments and is used to assess the quality of Generative Adversarial Networks samples. Fréchet distance between two Gaussians that have been fitted to feature representations of the Inception network, is used to calculate the FID Score. It is also essential to test the model with a minimum of 10 k images to obtain appropriate and truthful FID scores [48]. In this work, we tested on 11 k images to evaluate the FID SCORE. The best value of FID is obtained for $\lambda = 5$, as seen in Table 1. A lower FID score suggests greater image quality; however, not necessarily better realism since our images have better realism for $\lambda = 3$ (Fig. 10) but a higher FID score (Table 1).

The stop criteria for the GAN model is when it reaches the Nash Equilibrium. But since we typically employ the SGD, the loss of both G and D models oscillates and never reach Nash Equilibrium. So one of the better methods to stop the GAN training is by visually inspecting the generated images and early stop if there is no visually perceived improvement in the generated images. In this work, we applied the early stopping and observed that the FID scores for early stopping (450 Epochs) are better than late stopping (650 Epochs) for both $\lambda = 3$ and $\lambda = 5$ as shown in Table1.

Since the FID score cannot indicate whether the produced image from the captions is well conditioned on the text description provided. Therefore, we use R-precision, an evaluation metric for rating the retrieval performances, as an additional evaluation metric for the task of text-to-image synthesis. If there are R appropriate documents that are applicable to a query, we review the top 'R' ranked obtained results of the method and discover that 'r' is relevant, and therefore, R-precision is given by 'r/R'. We have performed a retrieval experiment in which we have use validation images to query the text that corresponds to them. To begin with, the global feature vectors of the output images and their text descriptions are extracted using the image and text encoders learned in DAMSM.

The next step is to compute the cosine similarity between the global image and global text vector and, lastly to calculate the R-precision. The candidate texts are ranked for every image in the order of descending similarity and select the top r valid descriptions. The model produces 11,000 photographs from randomly chosen unseen captions to calculate the R-precision. For each query picture, the candidate text descriptions consist of single ground truth (i.e., R = 1) and 99 randomly selected descriptions that don't fit. Table 1 shows the FID scores and R-precision achieved for different λ values.

3.4. Experimental setup for a standalone device

The birds and face trained model was deployed on a Raspberry Pi 4 Model B (4 GB RAM). The Raspberry Pi was interfaced with the VNC Viewer App. All the dependencies, PyTorch wheel file, pre-trained models and the code were put onto a 16 GB MicroSD Card and the evaluation code was executed from the command window.

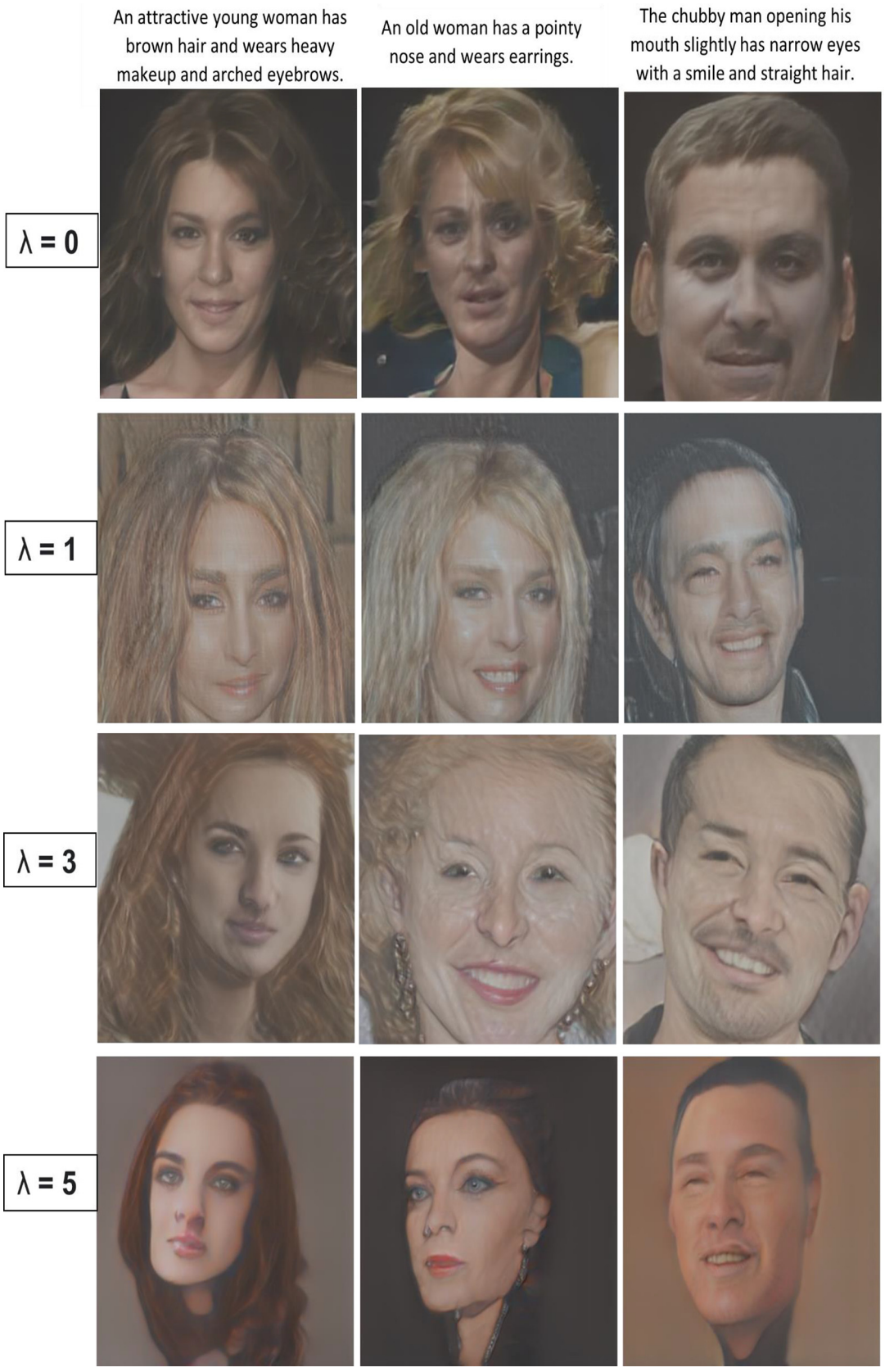

Fig. 10. Output image for given captions for different λ values.

Due to the restriction of RAM, it generated only three images from their corresponding captions. Any more input captions resulted in an 'Out of memory' error. The response time from the input to output was approximately 14 to 15 s. This testing on a standalone device was done to check how optimised and efficient the models are. Figs. 11 and 13 show the Input caption and Output image on

Table 1
The best FID score and the corresponding R precision rate of the AttnGAN model on CelebA dataset. More results in Fig. 10.

Method	FID SCORE	R-precision (%)
600 epochs		
AttnGAN2, λ = 0 No DAMSAM	53.11	11.71 ± 0.01
AttnGAN2, λ = 1	50.93	15.33 ± 0.01
AttnGAN2, λ = 3	56.41	26.83 ± 0.01
AttnGAN2, λ = 5	48.27	38.66 ± 0.01
Early stopping, 450 epochs		
λ = 3, 450 epochs	55.44	27.30 ± 0.0102
λ = 5, 450 epochs	40.73	39.60 ± 0.0203

the VNC viewer App. Figs. 12 and 14 show the total time for predicting the output on Raspberry Pi 4 Model B (4 GB RAM).

4. Discussion

4.1. Comparison

Table 2 shows existing work done in the fields of Text-to-image generation with the CelebA dataset. A number of approaches and methodologies have been proposed.

In [49,50], a multimodal CELEBA-HQ dataset is used. The dataset consists of 30 k high-resolution face images, each having a high-quality segmentation mask, sketch, and descriptive text. StyleGAN is used for face generation with a FID score reported as 106.37 and 101.42, respectively. In [41], DC GAN is used for face image generation with IS of 1.4 ± 0.7 and the limitations of using inception score as an evaluation metric for faces datasets are also discussed. In [51], a smaller subset of CelebA named as SCU-Text2face dataset is used. Two hundred samples are used for testing with a reported FID score of 44.49. However, the base FID paper [48] states that a minimum of 10 k testing images must be used for generating valid FID scores. As opposed to this, we have used 11 k testing images from CelebA dataset and obtained a FID score of 40.73 for $\lambda = 5$

4.2. Effect of semantic alteration on the images

The AttnGAN is not only capable of generating images of high resolution. In addition, it can also consider all the attributes mentioned in the input caption. Removing or replacing a certain keyword in the input drastically impacts the output image. An example of this can be seen in Figs. 15 and 16.

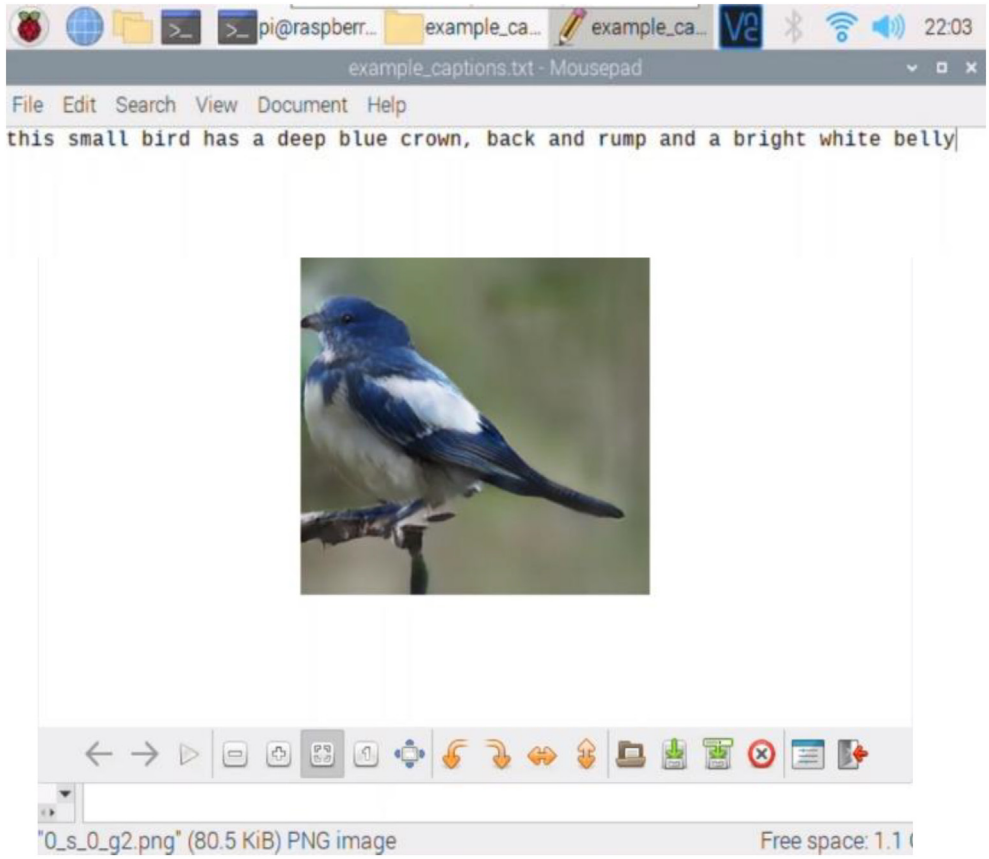

Fig. 11. Input caption and Output image for the birds dataset on the VNC viewer App.

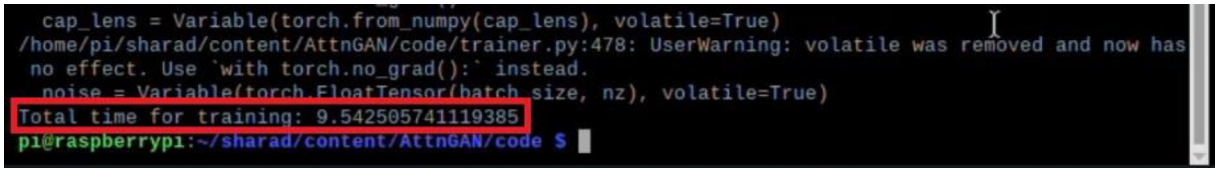

Fig. 12. Total time for predicting the output on Raspberry Pi 4 Model B (4 GB RAM) for birds data.

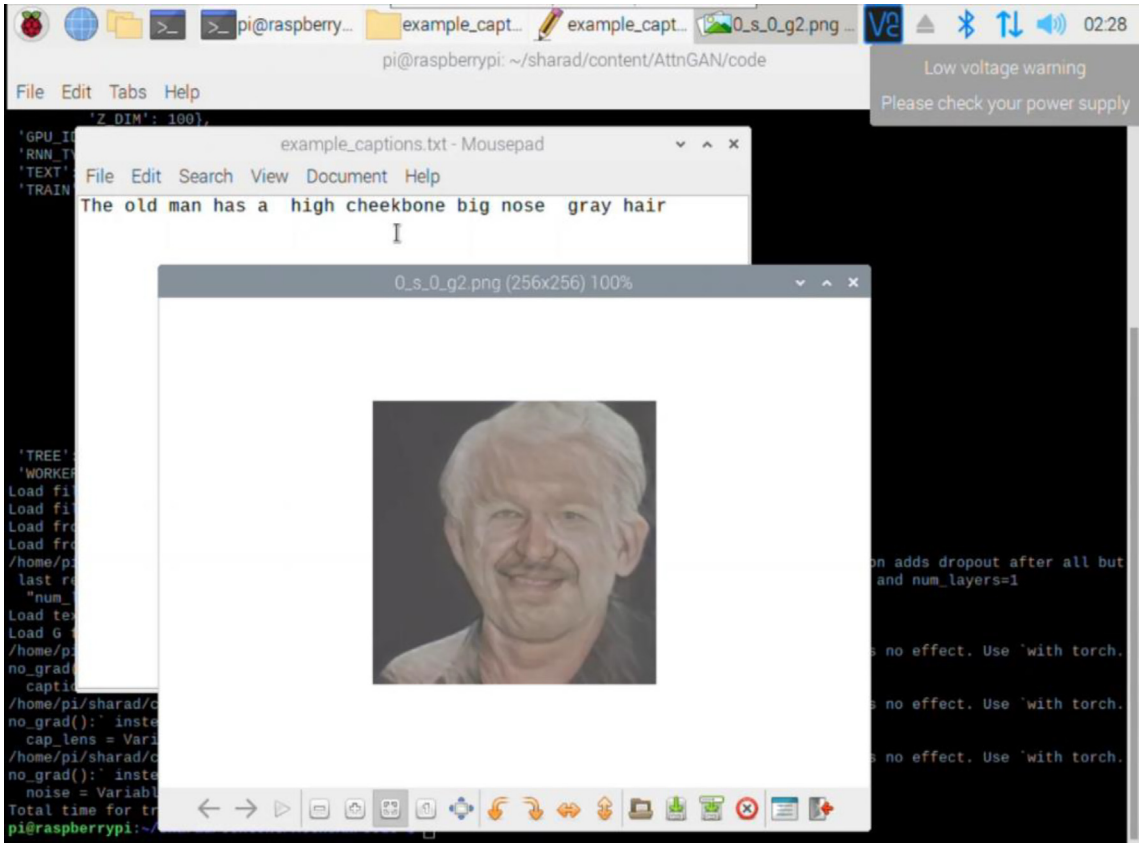

Fig. 13. Input caption and Output image for CelebA dataset on the VNC viewer App.

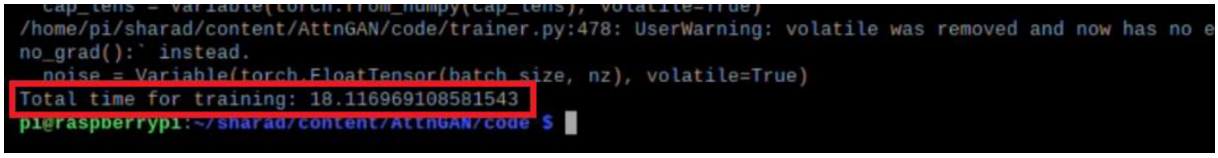

Fig. 14. Total time for predicting the output on Raspberry Pi 4 Model B (4 GB RAM) for CelebA data.

By altering some of the words in the input descriptions, we can observe how responsive the output images are to alterations in the input sentences. It shows how the generated visuals are altered in response to alterations in the input phrases, demonstrating that the model can detect even minor semantic alterations in the written description.

4.3. Challenges

4.3.1. Bias in the dataset

Both CelebA and Face2Text datasets are primarily focused on Caucasian ethnicity and have an over-representation of the same. However, it does not contain balance samples for fair and dark-skinned people leading to the under-representation of these ethnic groups. This is one of the limitations of these datasets [52] and to remedy this, a balanced dataset that has an equal and unbiased representation of ethnicity and gender must be developed. This kind of unbalanced dataset gives rise to unethical outcomes of the AI models.

4.3.2. Realism vs quality

We have observed that the AttnGAN occasionally produces photos that are clear and detailed. However, not necessarily realistic. FID score serves as a metric to evaluate the quality of the generated images in relation to the ground truth corresponding to the input textual description. It correlates well with the quality of the image. However, it does not necessarily represent the realism of the images. Therefore, the lower FID score, although is a metric of better image generation based on the textual description, does not necessarily mean the more realistic images. This is owing to the λ parameter and its effect on image generation. The qualitative analysis of generated images shown in Fig. 10 represents that for $\lambda = 3$ the images are more real-looking than for $\lambda = 5$. However, Table 1 exhibits that the FID score for $\lambda = 3$ is more than that of $\lambda = 5$. The balance between image quality vs realism is an important challenge of the generative models.

4.3.3. Use of better text encoding methods

The transformers are typically considered to perform better than LSTM, as reported in the literature [25]. However, there are a few approaches reported in recent times which suggest that transformers may not be the ultimate solution. In [53], authors propose that in the context of language models, convolutional models may prove competitive to Transformers when pre-trained. Also, in [54], it is suggested that replacing BERT with a linear transform such as Fourier transform proves to be exceptionally faster in real-time GPU implementations. Also, in [55], an approach

Table 2
Prominent work done in the fields of Text-to-image generation with the CelebA dataset.

	Dataset	Approach	Highest Output Image Resolution	Metrics
Xia et al. [49]	MULTI-MODAL CELEBA-HQ (has 30 k high-resolution face images, each having a high-quality segmentation mask, sketch, and descriptive text)	StyleGAN inversion module, visual-linguistic similarity learning, and instance-level optimisation.	1024x1024	FID: 106.37 Other Metrics used: LPIPS (Learned Perceptual Image Patch Similarity), Accuracy, Realism
Xia et.al. [50]	Multi-Modal CelebA-HQ ((has 30 k high-resolution face images, each having a high-quality segmentation mask, sketch, and descriptive text)	(Builds on TediGAN-A [49])	1024x1024	FID:101.42 Other Metrics used: LPIPS (Learned Perceptual Image Patch Similarity), Accuracy, Realism
Nasir et al. [41]	CelebA: 7500 training, 2500 testing	DC-GAN with GAN-CLS loss	64x64	IS: 1.4 ± 0.7
Chen et al. [51]	SCU-Text2face: 1000 training, 200 testing	FTGAN	256x256	FID: 44.49,Other Metrics used: IS, FSD, FSS
Ours	CelebA: 10.2 k samples, early stopping.	AttnGAN	256 × 256	FID: 40.73 (best with $\lambda = 5$ and early stopping after 450 epochs)

Fig. 15. The figure demonstrates the effect of specific words on the output generated. In the image, the word 'old' significantly affects how the face of the lady is generated. This demonstrates how AttnGAN can detect minor semantic alterations.

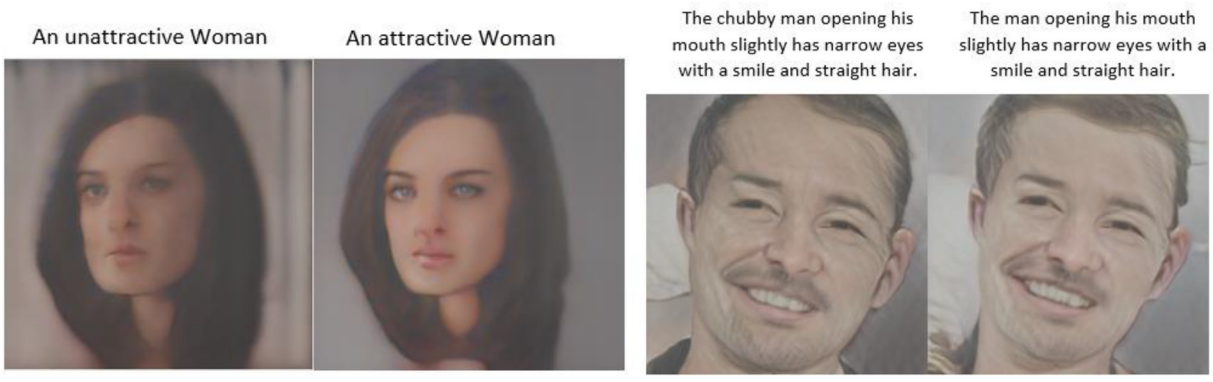

Fig. 16. Other examples showcasing how the words 'attractive' and 'chubby' are learned by the AttnGAN model.

based on auto-encoders with transformers is suggested for text-to-image generation.

In summary, there are multiple approaches to the task of text-to-image generation in general and text encoding in particular. We experimented with the simplest method of text encoding since the focus of this work is specifically on understanding the GAN models and experimenting with them extensively for face image generation. Our contribution lies in identifying various approaches of implementations of GANs for better realistic images, comparing and analysing the various evaluation and performance methods to strike a balance between realism and quality, optimization via DAMSM loss and most importantly, handling of the limitations posed by inconsistent human participation. Hence we chose to implement the text encodings via LSTM. However, we believe that replacing the LSTM with the transformer attention model will improve the performance, which will be the further extension of this work.

5. Conclusion

In general, Photorealistic image generation from its description is constrained on its dataset. Every word of the caption has an impact on the quality of the output image. In the case of text-to-face generation, if the dataset consists of more prior information on a face rather than focusing on selected attributes, it certainly increases the quality of the obtained results. Therefore, in this work, we proposed the implementation of text-to-face synthesis using AttnGAN. Initially, experiments were conducted on StackGAN and PGGAN for the Birds and Flowers dataset. But owing to the lack of performance of these architectures on more complex datasets and lack of focusing attention to a specific attribute, AttnGAN was employed. AttnGAN was used on the Face2Text dataset and CelebA dataset.

The model was first implemented on the Face2Text dataset. Following this, we trained the model on 10.2 k images from the CelebA dataset. DAMSM loss was considered for optimisation, and we experimented with variousλvalues. The results obtained by our model are compared with the existing models employed on the CelebA dataset. Our model outperforms the other approaches in terms of using the required number of testing samples and generating the lowest FID scores. We also studied and demonstrated the effect of semantic alterations on the generated images. Such images are very similar to each other. However, they have distinct variations introduced due to semantic alterations. The effect of FID and λ values on the quality and realism of the image is analysed, and an early stopping method is implemented to achieve the balance between the same. Certain challenges, specifically due to the bias in the datasets, are also discussed. Finally, we deployed these trained models of the birds and faces dataset on a Raspberry Pi to achieve real-world usability, accessibility and portability of this framework. Deploying the model as an API has enormous promise in the field of public safety and increased useability. Future work may focus on capturing global coherent structures as well as employing the attention trasnfromers model for advanced text encoding.

CRediT authorship contribution statement

Sharad Pande: Software, Validation, Writing – original draft. **Srishti Chouhan:** Software, Writing – original draft, Visualization, Validation. **Ritesh Sonavane:** Software, Writing – original draft. **Rahee Walambe:** Conceptualization, Methodology, Wiriting - original draft, Supervision. **George Ghinea:** Review and Suggestions. **Ketan Kotecha:** Conceptualization, Methodology, Supervision, Validation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee, in *33rd International Conference on Machine Learning*, 2016, 2016,.

[2] R. Zhou, C. Jiang, Q. Xu, A survey on generative adversarial network-based text-to-image synthesis, Neurocomputing 451 (2021) 316–336, https://doi.org/10.1016/j.neucom.2021.04.069.

[3] T. Xu et al., "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks," 2018. doi: 10.1109/CVPR.2018.00143.

[4] N. Zeng, H. Li, Z. Wang, W. Liu, S. Liu, F.E. Alsaadi, X. Liu, Deep-reinforcement-learning-based images segmentation for quantitative analysis of gold immunochromatographic strip, Neurocomputing 425 (2021) 173–180, https://doi.org/10.1016/j.neucom.2020.04.001.

[5] N. Zeng, Z. Wang, H. Zhang, K.-E. Kim, Y. Li, X. Liu, An Improved Particle Filter with a Novel Hybrid Proposal Distribution for Quantitative Analysis of Gold Immunochromatographic Strips, IEEE Transactions on Nanotechnology 18 (2019) 819–829, https://doi.org/10.1109/TNANO.772910.1109/TNANO.2019.2932271.

[6] N. Zeng et al., "Image-based quantitative analysis of gold immunochromatographic strip via cellular neural network approach," IEEE Transactions on Medical Imaging, vol. 33, no. 5, 2014, doi: 10.1109/TMI.2014.2305394.

[7] I. J. Goodfellow et al., "Generative Adversarial Networks," Communications of the ACM, vol. 63, no. 11, pp. 139–144, Jun. 2014, Accessed: Jul. 14, 2021. [Online]. Available: https://arxiv.org/abs/1406.2661v1

[8] H. Zhang et al., "StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks," in Proceedings of the IEEE International Conference on Computer Vision, 2017, vol. 2017-October. doi: 10.1109/ICCV.2017.629.

[9] S. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, "Learning what and where to draw," 2016.

[10] H. Zhang et al., "StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 8, 2019, doi: 10.1109/TPAMI.2018.2856256.

[11] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive Growing of GANs for Improved Quality, Stability, and Variation," 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings, Oct. 2017, Accessed: Jul. 14, 2021. [Online]. Available: https://arxiv.org/abs/1710.10196v3

[12] T. Y. Lin et al., "Microsoft COCO: Common objects in context," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2014, vol. 8693 LNCS, no. PART 5. doi: 10.1007/978-3-319-10602-1_48.

[13] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep Learning Face Attributes in the Wild." pp. 3730–3738, 2015. Accessed: Jul. 14, 2021. [Online]. Available: http://personal.ie.cuhk.edu.hk/

[14] M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," Nov. 2014, Accessed: Jul. 14, 2021. [Online]. Available: https://arxiv.org/abs/1411.1784v1

[15] A. Radford, L. Metz, and S. Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," 4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings, Nov. 2015, [Online]. Available: https://arxiv.org/abs/1511.06434v2

[16] H. Dong, S. Yu, C. Wu, and Y. Guo, "Semantic Image Synthesis via Adversarial Learning," in Proceedings of the IEEE International Conference on Computer Vision, 2017, vol. 2017-October. doi: 10.1109/ICCV.2017.608.

[17] T. Qiao, J. Zhang, D. Xu, D. Tao, in: in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019, p. 2019-June., https://doi.org/10.1109/CVPR.2019.00160.

[18] Z. Zhang, Y. Xie, and L. Yang, "Photographic Text-to-Image Synthesis with a Hierarchically-Nested Adversarial Network," 2018. doi: 10.1109/CVPR.2018.00649.

[19] S. Hong, D. Yang, J. Choi, and H. Lee, "Inferring Semantic Layout for Hierarchical Text-to-Image Synthesis," 2018. doi: 10.1109/CVPR.2018.00833.

[20] S. Sharma, D. Suhubdy, V. Michalski, S. E. Kahou, and Y. Bengio, "ChatPainter: Improving text to image generation using dialogue," 2018.

[21] H. Dong, J. Zhang, D. Mcllwraith, and Y. Guo, "I2T2I: Learning text to image synthesis with textual data augmentation," in Proceedings - International Conference on Image Processing, ICIP, 2018, vol. 2017-September. doi: 10.1109/ICIP.2017.8296635.

[22] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016, vol. 2016-December. doi: 10.1109/CVPR.2016.10.

[23] K. Xu et al., "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," in Proceedings of the 32nd International Conference on Machine Learning, Jul. 2015, vol. 3.

[24] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-Attention Generative Adversarial Networks," in Proceedings of the 36th International Conference on Machine Learning, Jul. 2019, vol. 2019-June.
[25] A. Vaswani et al., "Attention Is All You Need," Advances in Neural Information Processing Systems, vol. 2017-December, Jun. 2017.
[26] L. Ye, B. Zhang, M. Yang, W. Lian, Triple-translation GAN with multi-layer sparse representation for face image synthesis, Neurocomputing 358 (2019) 294–308, https://doi.org/10.1016/j.neucom.2019.04.074.
[27] A. Gatt et al., "Face2Text: Collecting an annotated image description corpus for the generation of rich face descriptions," 2019.
[28] J. He, J. Zheng, Y. Shen, Y. Guo, H. Zhou, Facial Image Synthesis and Super-Resolution With Stacked Generative Adversarial Network, Neurocomputing 402 (2020) 359–365, https://doi.org/10.1016/j.neucom.2020.03.107.
[29] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," in Proceedings of the IEEE International Conference on Computer Vision, 2017, vol. 2017-October. doi: 10.1109/ICCV.2017.244.
[30] A. Brock, J. Donahue, and K. Simonyan, "Large Scale GAN Training for High Fidelity Natural Image Synthesis," 7th International Conference on Learning Representations, ICLR 2019, Sep. 2019.
[31] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2019, vol. 2019-June. doi: 10.1109/CVPR.2019.00453.
[32] Y. Choi, M. Choi, M. Kim, J. W. Ha, S. Kim, and J. Choo, "StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation," 2018. doi: 10.1109/CVPR.2018.00916.
[33] T. C. Wang, M. Y. Liu, J. Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs," 2018. doi: 10.1109/CVPR.2018.00917.
[34] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua, "Towards Open-Set Identity Preserving Face Synthesis," 2018. doi: 10.1109/CVPR.2018.00702.
[35] R. Huang, S. Zhang, T. Li, and R. He, "Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis," Proceedings of the IEEE International Conference on Computer Vision, vol. 2017-October, 2017, doi: 10.1109/ICCV.2017.267.
[36] X. Chen, L. Qing, X. He, J. Su, Y. Peng, From Eyes to Face Synthesis: A New Approach for Human-Centered Smart Surveillance, IEEE Access 6 (2018) 14567–14575, https://doi.org/10.1109/ACCESS.2018.2803787.
[37] X. Di and V. M. Patel, "Face synthesis from visual attributes via sketch using conditional vaes and gans," arXiv preprint arXiv:1801.00077, 2017.
[38] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, T.S. Huang, "Generative Image Inpainting with Contextual Attention" (2018), https://doi.org/10.1109/CVPR.2018.00577.
[39] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," 2011.
[40] M. E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," 2008. doi: 10.1109/ICVGIP.2008.47.
[41] O. R. Nasir, S. K. Jha, M. S. Grover, Y. Yu, A. Kumar, and R. R. Shah, "Text2FaceGAN: Face generation from fine grained textual descriptions," 2019. doi: 10.1109/BigMM.2019.00-42.
[42] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016, vol. 2016-December. doi: 10.1109/CVPR.2016.308.
[43] O. Russakovsky et al., "ImageNet Large Scale Visual Recognition Challenge," International Journal of Computer Vision, vol. 115, no. 3, 2015, doi: 10.1007/s11263-015-0816-y.
[44] C. Bodnar, "Text to Image Synthesis Using Generative Adversarial Networks," May 2018, doi: 10.13140/rg.2.2.35817.39523.
[45] Karnewar Animesh and Ibrahim Ahmed Hani, "GitHub - akanimax/T2F: T2F: text to face generation using Deep Learning." https://github.com/akanimax/T2F (accessed Jun. 01, 2021).
[46] "GitHub - 2KangHo/AttnGAN-CelebA: Face Image Generation using AttnGAN with CelebA Dataset." https://github.com/2KangHo/AttnGAN-CelebA (accessed Jun. 01, 2021).
[47] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in Advances in Neural Information Processing Systems, 2017, vol. 2017-December.
[48] "GitHub - bioinf-jku/TTUR: Two time-scale update rule for training GANs." https://github.com/bioinf-jku/TTUR (accessed Jun. 01, 2021).
[49] W. Xia, Y. Yang, J.-H. Xue, and B. Wu, "TediGAN: Text-Guided Diverse Face Image Generation and Manipulation," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2256-2265, Dec. 2020, [Online]. Available: https://github.com/weihaox/TediGAN.
[50] W. Xia, Y. Yang, J.-H. Xue, and B. Wu, "Towards Open-World Text-Guided Face Image Generation and Manipulation," Apr. 2021.
[51] 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.
[52] E. M. Rudd, M. Günther, and T. E. Boult, "MOON: A mixed objective optimization network for the recognition of facial attributes," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2016, vol. 9909 LNCS. doi: 10.1007/978-3-319-46454-1_2.
[53] Y. Tay et al., "Are Pre-trained Convolutions Better than Pre-trained Transformers?," arXiv preprint arXiv:2105.03322, 2021.
[54] "Google Replaces BERT Self-Attention with Fourier Transform: 92% Accuracy, 7 Times Faster on GPUs | Synced." https://syncedreview.com/2021/05/14/deepmind-podracer-tpu-based-rl-frameworks-deliver-exceptional-performance-at-low-cost-19/amp/ (accessed Jul. 14, 2021).
[55] N. A. Fotedar and J. H. Wang, "Bumblebee: Text-to-Image Generation with Transformers", Accessed: Jul. 14, 2021. [Online]. Available: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/reports/custom/15709283.pdf

Sharad Pande is a undergraduate student at Symbiosis Institute of Technology. He is pursuing the Bachelors of Technology, majoring in Electronics and Telecommunication. He has keen interest in machine learning and data science. For the past few years he us working in the area of GANs and their use for various generative tasks.

Srishti Chouhan is a undergraduate student at Symbiosis Institute of Technology. She is pursuing the Bachelors of Technology, majoring in Electronics and Telecommunication. She has keen interest in deep learning and data analysis. For the past few years she is working in the area of image processing, GANS and deep learning methods for various applications.

RItesh Sonavane is a undergraduate student at Symbiosis Institute of Technology. He is pursuing the Bachelors of Technology, majoring in Electronics and Telecommunication. He has keen interest in robotics and implementation of various models on hardware platforms. For the past few years he is working on deployment of various models on microprocessors and hardware platforms.

Rahee walambe received MPhil, Ph.D. Degree from Lancaster University, UK, in 2008. From 2008 to 2017, she was a research Consultant with various organizations in the control and robotics domain. Since 2017, she has been working as an Associate Professor at Dept of Electronics and Telecommunications at Symbiosis Institute of Technology, Symbiosis International University, Pune, India. Her area of research is applied Deep Learning and AI in the field of Robotics and Healthcare. She is awarded number of national and international research grants.

George Ghinea is a Professor in the Department of Computer Science at Brunel University London. My research activities lie at the confluence of Computer Science, Media and Psychology. He has applied my expertise in areas such as eye-tracking, telemedicine, multi-modal interaction, and ubiquitous and mobile computing. I am particularly interested in building human-centred e-systems, particularly integrating human perceptual requirements. His work has been funded by both national and international funding bodies.

Ketan kotecha pursued Ph.D.& MTech from (IIT Bombay) and is currently holding the positions as Head, Symbiosis Centre for Applied AI (SCAAI), Director, Symbiosis Institute of Technology, Dean, Faculty of Engineering, Symbiosis International (Deemed University). He is an expert in AI and Deep Learning. He has published 100+ widely in a number of excellent peer-reviewed journals on various topics ranging from cutting-edge AI, education policies, teaching-learning practices and AI for all. He has published 3 patents and delivered keynote speeches at various national and international forums. He is a recipient of multiple international research grants and awards.