IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 23, 2020, accepted July 28, 2020, date of publication August 10, 2020, date of current version January 5, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3015656

A Realistic Image Generation of Face From Text Description Using the Fully Trained Generative Adversarial Networks

MUHAMMAD ZEESHAN KHAN[1], SAIRA JABEEN[1], MUHAMMAD USMAN GHANI KHAN[2], TANZILA SABA[3], (Senior Member, IEEE), ASIM REHMAT[2], AMJAD REHMAN[3], (Senior Member, IEEE), AND USMAN TARIQ[4]

[1]Alkhawarizmi Institute of Computer Sciences, UET Lahore, Lahore 54000, Pakistan
[2]Department of Computer Science and Engineering, UET Lahore, Lahore 54000, Pakistan
[3]Artificial Intelligence and Data Analytics Lab, CCIS Prince Sultan University, Riyadh 11586, Saudi Arabia
[4]College of Computer Engineering and Science, Prince Sattam bin Abdulaziz University, Alkharj 16278, Saudi Arabia

Corresponding author: Amjad Rehman (drrehman70@gmail.com) and Usman Tariq (u.tariq@psau.edu.sa)

This work was supported in part by the National Center of Artificial Intelligence, Pakistan. and in part by the Artificial Intelligence and Data Analytics Lab, Prince Sultan University, Riyadh, Saudi Arabia.

ABSTRACT Text to face generation is a sub-domain of text to image synthesis. It has a huge impact on new research areas along with the wide range of applications in the public safety domain. Due to the lack of dataset, the research work focused on the text to face generation is very limited. Most of the work for text to face generation until now is based on the partially trained generative adversarial networks, in which the pre-trained text encoder has been used to extract the semantic features of the input sentence. Later, these semantic features have been utilized to train the image decoder. In this research work, we propose a fully trained generative adversarial network to generate realistic and natural images. The proposed work trained the text encoder as well as the image decoder at the same time to generate more accurate and efficient results. In addition to the proposed methodology, another contribution is to generate the dataset by the amalgamation of LFW, CelebA and locally prepared dataset. The dataset has also been labeled according to our defined classes. Through performing different kinds of experiments, it has been proved that our proposed fully trained GAN outperformed by generating good quality images by the input sentence. Moreover, the visual results have also strengthened our experiments by generating the face images according to the given query.

INDEX TERMS GAN, CNN, text to face, image generation, face synthesis, data augmentation, legal identity for all.

I. INTRODUCTION

Generating images using the text description is one of most challenging and important tasks in machine learning. This task involves handling the language modalities problems which include the control and management of incomplete and ambiguous information using the natural language processing techniques and algorithms. After that, this information is used to learn by computer vision approaches and algorithms Currently, it is one of the latest research domains in computer vision.

Generating images from text is the opposite process of image captioning and image classification, where text and caption are generated from images. Just like the image captions, text to image generation helps to find context and relationship between the image and the text along with exploring human visual semantics. Moreover, it has a large number of applications in art, designs, image retrieval and searching. Currently, most of the methods for generating images from the text are based on the traditional method in which the pre-trained text encoder has been utilized to get the semantic vector from input descriptions. Based on the semantic vectors, conditional GAN is trained to generate realistic face images. Although this method generates high-quality face images, they split the training method into two steps; train the

The associate editor coordinating the review of this manuscript and approving it for publication was Pengcheng Liu.

text encoder and image decoder separately. Most of the generative adversarial networks focus on the generation of the synthesized images using the sentence level information. Generating the images using the sentence level information probably has chances of information loss at word level. As a result, the accurate images cannot be generated [1], [2]. Most of the work which was done for the problem of " Text to Image" generation is based on simple dataset problems such as birds [3] and flowers [4]. However, the work that mapped the objects along with scenes was very limited. To overcome this problem, [2] utilized the AttnGAN, they were failed to achieve good results as their output image was semantically not meaningful. They tried to explore the COCO dataset and mapped the object along with the scene with the sentence-level information. However, the object and word-level information was still missing.

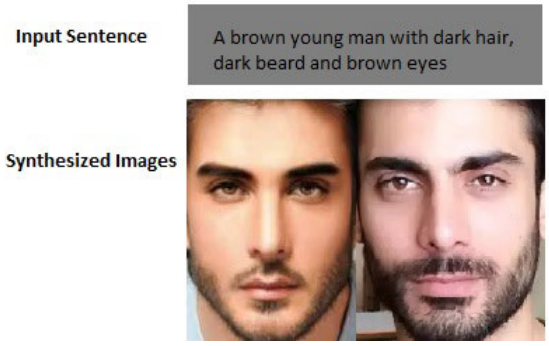


FIGURE 1. Two images for text to image synthesis system that is referenced with same input sentence.

Text to face image generation is the subdomain of the text to image generation, where the ultimate goal is to generate the image using the user-specified description about the face. So, there are two major tasks of generating face images from text. Fig. 1 shows the input and output for a text to image synthesis system. It can be observed that text to face synthesis involves generating high-quality images and generating the appropriate images related to the given description. This task of generating the face images from the text description is more relevant to the public safety tasks. For example, we consider the scenario of the crime scene. In most of the cases, the witness of the crime scene has appeared before the law enforcement agencies to help in drawing the portrait of the suspected criminal. The witness tells the description of the criminal to the portrait maker, then he/she draws the portrait of the criminal on the drawing board. The proposed work will help to automate the whole task by negating the role of the portrait maker. The manual work is tedious and time-consuming and requires professional knowledge and experience. Thus, this work will be helpful for law-enforcement agencies.

There are different datasets available for text to image syntheses, like CUB [5], Oxford102 [6], and COCO [7]. But there is no standard dataset, which is available for text to face generation. The work is done in the domain of text to face generation is very limited. In this paper, a self-generated dataset is also presented with the help of the google image search and two publically available datasets for face to text generation.

The main motivation behind this research work is to generate the synthesized images of the face based on the text description. The proposed algorithm in this paper has ensured to generate high-quality images by preserving the face identity. Moreover, it is also capable of generating the exact images based on the given descriptions. This research work has also been utilized in many industrial applications like automatic sketch making of the suspected face in crime investigation departments.

We have made the following contributions in our paper.

1. Generating the dataset related to the text to face images.
2. Proposed a Fully Trained Generative Adversarial Network, which has a trainable text encoder as well as a trainable image decoder.
3. Two discriminators are proposed to utilize the strength of joint learning.
4. Generating the photo-realistic images of the faces from the description by preserving details.

The rest of the paper is organized as follows; Literature survey has been discussed in section II, proposed methodology and framework design have been briefly discussed in section III. Whereas, dataset description and experimental analysis described in IV and V, respectively. Paper has been concluded in Section VI.

II. LITERATURE SURVEY

Two domains are related to the work. The first one is the text to image synthesis and the second one is the text attributes to face generation. Both domains are discussed one by one as follow;

A. TEXT TO IMAGE GENERATION

There are a lot of frameworks available for the text to image generation. These frameworks are based on the encoding through the encoder and decoding through decoder also by using the conditional GAN. The text is encoded through an encoder that processes sequential information of text and the image is decoded using the spatial decoder. The text encoder encodes the input description into the semantic vectors, whereas the image decoder uses these semantic vectors to generate the natural and realistic images. There are two basic purposes of the text to image synthesis, the first one is to generate the natural and realistic images and the second one is to make sure that generated images are related to the given description. All basic algorithms and procedures for text to image generation are based on this rule of thumb.

From the past few years, the work on the generative network has been boasted up for image synthesis. Kingma et.al [8] utilized the stochastic backpropagation to train the auto-variational encoder for data generation purposes. Since the birth of the generative adversarial network, which was

proposed by Goodfellow *et al.* [1] researchers have studied and researched it widely. The very first task which focused the text to image generation has been done by Reed et. al. [9]. They have utilized the conditional GAN to build the two end to end network for text to image generation. They have obtained the semantic vectors from the text by using the pre-trained Char-CNN-RNN and used these vectors to decode the natural images using the decoder which is very similar to DCGAN [10].

After that, researchers have started to make further progress in this particular domain [11]. Zhang et al [12] proposed the StackGAN, which is based on two stages and generates high-quality images with the improved inception score. Till then, researchers were able to generate high-quality images. The focus at that time shifted to improve the similarity between the text and images. Reed et.al [13] proposed a network that generates images based on the first generated box. This produced more efficient and accurate results on the output images. Sharma et.al. [14] introduced the mechanism of dialogue to enhance the understanding of the text. They claimed that the method helped them to achieve good results for the image synthesis relevant to the input text. Dong *et al.* [11] proposed and introduced a new approach for the image to image and text to image generation. Moreover, they also introduced the training mechanism of image-text-image. They first generated the text from the images, and then this text was used to generate the images.

The attention-based mechanism has gained a lot of success in the image and text-related tasks. Researchers have also utilized the attention mechanism in generating text to image task. Xu *et al.* [15] first utilized the attention mechanism to generate the images from the text. They have introduced the AttnGAN to generate high-quality images from the text by applying natural language processing techniques and algorithms. Qiao *et al.* [16] proposed the approach, which was based on the global-local collaborative attention model. Zhang et.al [17] proposed an approach that was based on visual semantic similarity. So, as a conclusion, we can say that these researchers currently have focused to boost up the consistency between the generated images and input text.

B. TEXT TO FACE GENERATION

Since the invention of the GAN, which was proposed by Goodfellow [1] in 2014, image synthesis using deep learning techniques become the hot topic of research [18]. There are two large scale datasets which are publically available for face synthesis task. These datasets are the CelebA [19] and LFW [20]. Face synthesis is very popular among the research community. Most of the state of the artwork has tested their model capabilities and abilities for face synthesis using the GAN and conditional GAN. DCGAN [21], CycleGAN [22], Pro-GAN [6], BigGAN [16], StyleGAN [10], StarGAN [9] are the examples of this problem. The quality of the generated face images is improving day by day with the development in the generative adversarial networks.

Some of the networks can generate good quality face images with a size of 1024×1024. These face images are much larger than the original images present in the face dataset. These described models first learned through the noise vector with the help of mapping and followed the normal distribution to generate the natural images of the face. However, they are not able to generate an accurate and precise face based on the input description.

To overcome and tackle this problem, many researchers have worked on different directions of face synthesis. These directions include converting the face edges into the natural face images [23], swapping the facial attributes of two different face images [24], generating the face with the help of the side face [25], generating the face with the help of the human eye's region [26], draw sketches from the human face [27], face make-up [28] and many more. But as per our best knowledge, no one combined the different face-related information in a single methodology to generate the natural and realistic face images.

Some of the researchers have also worked on the face generation through the attributes description. Li et. al [32] proposed the work, in which they generated the face with the help of the attribute description by making sure that they preserve the identity of the face. The drawback of their proposed methodology is that it is only applicable to those faces which can be generated using the simple attributes. Another work named TP-GAN [33] has been proposed by the researchers. In this work, they have proposed the generative adversarial network based on the two pathways. They synthesized the frontal face images using the proposed network. Although they succeeded to generate the good results but required a large amount of labeled data of frontal faces. Some of the researchers have also explored the disentangled representation learning for face synthesis using the defined attributes of the face. DC-IGN [34] has proposed the variational auto-encoder using the patterns and techniques of disentangled representation learning. However, the major drawback of this work is that it only tackles one attribute in particularly one batch. It makes it computationally weak as well as it also requires the large explicitly annotated data for training. Luan et. al [35] proposed the algorithm, which they named as the DR-GAN. It is used for the learning purpose of generative and discriminative representation of face synthesis. Their proposed work was based on the poses of the face and did not focused on specified face attributes. However, our proposed framework makes sure to preserve the identity of the generated image by incorporating all the attributes information related to the face.

As per our best knowledge and based on the literature survey, the work on the face generation through the attribute description using the generative adversarial network is very less. Most of the work on this problem is done on the limited scope and failed to generate impressive results by not preserving the face identity. Moreover, most of the relevant proposed networks have trained the image decoder and used the

pre-trained text encoder. So, in this work we have proposed the fully trainable generative adversarial network.

Research gap matrix has also been shown in the Table 1.

TABLE 1. Research matrix of text to face.

	Generated faces based on			
Paper	Edge Information	Face Attribute	Eyes Information	Sketch Description
T.Wang et.al [23]	✓	×	×	×
J.Bao et.al [24]	×	✓	×	×
X.Chen, et.al [26]	×	×	✓	×
X. Di et.al [27]	×	×	×	✓
Proposed	✓	✓	✓	✓

III. PROPOSED METHODOLOGY AND FRAMEWORK DESIGN

In this methodology section, we will describe our proposed work in detail. Our proposed work is presented in two sections. In the first section, we discussed, how to encode the text into the semantic vectors, whereas in the second section we describe the mechanism of decoding the semantic features of text into the realistic natural images. A comprehensive overview of the whole network architecture is also described in detail. The framework design of our proposed architecture is based on two streams. In the first portion, the text is encoded, whereas in the second part image is decoded using the encoded text embeddings. A text encoder converts the textual data into a semantic vector. After that, the image decoder generates realistic images using the semantic features of the text, which is encoded by the text encoder. Currently, most of the text to image generation techniques of the generative adversarial network is based on the training of the separate modules. They trained the text encoder and image decoder separately. They have utilized the pre-trained text encoder or fully trained the text encoder for the training purpose of the image encoder. Whereas, in our proposed work, the entire framework is trained with text encoder and image encoder at the same time. The design of our fully trained generative adversarial network has been shown in Fig. 2. The training mechanism has also been incorporated on the text encoder to generate natural and realistic images. As shown in Fig. 2, our proposed architecture based on the auto-encoder and decoder framework. The encoder firstly encodes the sequences of input sentences to the semantic vector and using these vectors, natural images are generated. Both of these tasks have equal weights and importance in the text to image synthesis. The reason behind not utilizing the pre-trained text encoder because it has a direct link with the upper limits of image decoder and it affects the accuracy to generate the quality images. The main task of text to image synthesis is to generate high-quality images which is relevant and according to the input sentence. Utilizing the pre-trained text-encoder can generate realistic images to some extent. However, we cannot confirm that generated images are according to the input text without human evaluation or realistic. Because the images are generated based on the semantic vectors, which have been extracted using the pre-trained text encoder. So, to generate good results, training of both text decoder and image encoder is performed concurrently.

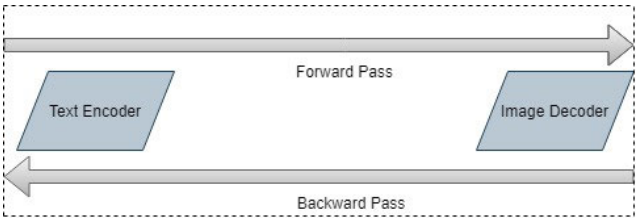


FIGURE 2. Design of our fully trained generative adversarial network.

The Architecture of Our Proposed Fully Trained Generative Adversarial Network: In this section, we describe the details of our proposed fully trained generative adversarial network. Fig. 2 depicts the pictorial representation of our network architecture. From Fig. 4 and Fig. 5, it clearly states that our proposed architecture of conditional GAN is based on one generator and two discriminators. The generator contains the trainable encoder and decoder, and it is the backbone of our proposed architecture for text to image task. It has been divided into two parts; first one is the text encoder and the second one is the image decoder. The details of these sections are as follows;

A. TEXT ENCODER

In our proposed work, text encodings have been extracted using the bidirectional LSTM as shown in Fig. 3. By using the bidirectional LSTM, we have extracted the semantic features from the given input text sentence.

In the proposed bidirectional LSTM, each word present in the sentence is connected with the two hidden states. Each hidden state corresponds to the one direction. The outputs of these two hidden states have been concatenated to get the semantic meaning against each word of the sentence. The input sentence is encoded in the form of the matrix using equation 1.

$$e \in \mathbb{R}^{D_x T} \tag{1}$$

Here in equation 1, T represents the total number of words along with the D dimensions. Whereas, e represents the feature vector for the i^{th} word present in the sentence. Firstly, each word embedding has been extracted to get the semantic features, which later become the input of the image generation process. The input to the image decoder for the image generation process is not in the form of a single word embedding of the sentence. So, to feed the image decoder network, the outputs of the last hidden states of the bidirectional LSTM have been concatenated and passed to the image decoder network.

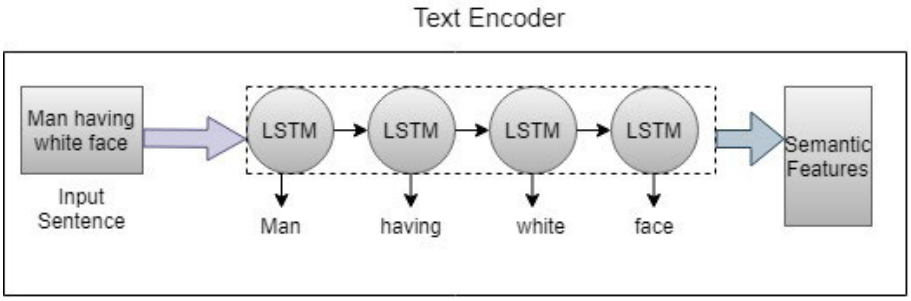


FIGURE 3. Text encoder.

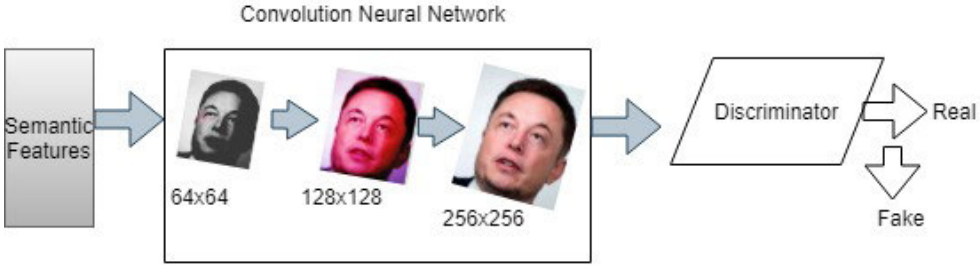


FIGURE 4. Image decoder.

The states of Bidirectional LSTM have been concatenated on the global sentence vector using the following equation 2.

$$C \in \mathbb{R}^D \tag{2}$$

Here in equation 2, C represents the semantic vector, which is concatenated with some noise having some dimensions D to generate the new vector. These generated vectors are fed as input to the image decoder network.

B. IMAGE DECODER

Our proposed convolution neural network is based on the three blocks as depicted in Fig. 4. Each block contains the 3 deconvolution layers. So, we have total 3 blocks and 9 deconvolution layers, which upsample feature map twice to its original size. The layers present into the blocks take the input from the encoded features of text as semantic vectors and generates realistic images. In the first stage, the semantic vectors are extracted from the text along with the noise concatenation and are passed as an input, and then these input vector is reduced to the 4×4 feature map. In all blocks, deconvolution has been performed on the feature maps. The up-sampling on the feature map increases the size twice the feature maps in all three layers of each block. So, the size of the feature map is up-sampled to 8×8, 16×16 and 32×32. After performing all the operations, feature maps are passed to the fully connected layers. In the second and third block, a similar task of up-sampling has been performed. There is a fine-tune block between the first, second and third blocks. Fine-tuning block contains the 3×3 kernel. The up-sampling block helped in fine-tuning the training parameters. The input to the second block is the feature map with a size of 64×64. The second block contains the same deconvolution layers the same as the first block and outputs the 128×128 feature map.

Whereas, in the 3rd block, we get the 256×256 feature map using the same layer architecture which was previously used in the first two blocks. After the three blocks of up-sampling layers, we have generated the 256×256 image, which later is used to calculate the generator loss.

TABLE 2. Generator architecture.

Block	First		Second	Third
Input Features	4x4		64x64	128x128
Number of De-Convolutional layers	3		1	1
Filters Size	3x3		3x3	3x3
Deconvolution layers output	**1**	8x8	128x128	256x156
	2	16x16		
	3	32x32		
	4	64x64		
Output Features	64x64		128x128	256x256

Table 2 shows the architectural detail of our proposed generator network. It specifies the details of input features, deconvolution layers, filter size, the output from deconvolution layers as well as the output features from the defined

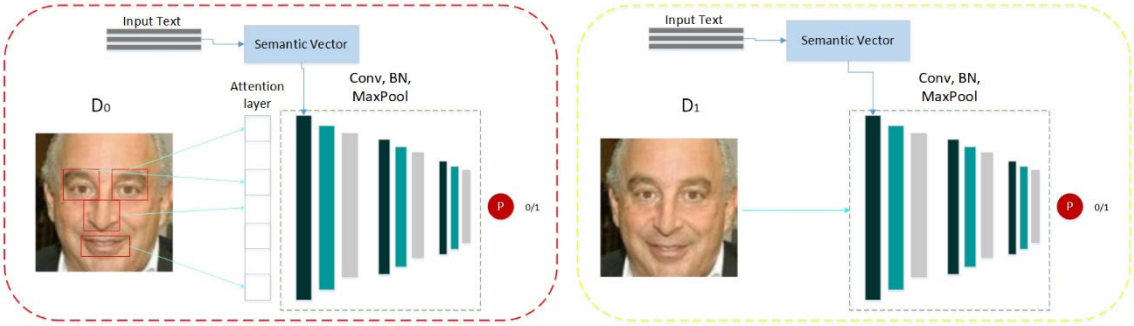


FIGURE 5. Discriminator networks.

three blocks of the architecture. This output is further passed to the discriminator network, to find the effectiveness of the generated face features.

C. DISCRIMINATOR

We propose a discriminator that measures the realness of human face region as well as the face features. A generated image along with its sentence encoding is passed to the CNN network that extracts the low-level region features using attention mechanism to be compared with ground truth image. The attention layer is added that allows convolution layers of discriminator to attend the region based features of eyes, nose, lips as well as entire facial features. Two stream discriminator network is designed as shown in Fig. 5.

The red box explains the attention-based discriminator D_0 network and the yellow box explains another discriminator network D_1. In discriminator D_0, an attention vector is initialized to focus on eye regions, lips region and nose region. Semantic vector representation from original sentence corresponding to these features is concatenated to attention features and then passed to succeeding convolutional layers.

Finally, the unconditional probabilities are computed to determine the correctness of local facial features. There are three convolutional layers combined with batch normalization layer and max-pooling layer. These layers minimize the feature representation of facial image with size $64 \times 64 \times 3$. Each convolution layer applies a convolution of filter size 4×4. Also, the max-pooling layer applies 2×2 sized filter to pool the strong weights. The semantic sentence vector is passed to second convolutional layer set. Since the vector representation of entire face features is also computed to measure the consonance of local features. Three sets of convolutional layers are adopted here with a convolution filter of size 4×4 and max pool filter of size 2×2. Discriminator D_1 has the same architecture as the D_0 without the attention layer. Final loss can be computed as a sum of two losses that are measured in an adversarial manner.

$$loss_{total} = loss_{D_0} + loss_{D_1} \tag{3}$$

where in equation 3, $loss_{D_0}$ is the cross-entropy loss for D_0 and $loss_{D_1}$ is for D_1. These losses are computed based on the unconditional probabilities p_0^t and p_1^t at the output neuron of D_0 and D_1, respectively.

$$p_i^t(\acute{z}_i) = \frac{e^{\acute{z}_i}}{\sum_{n=1}^{N} e^{\acute{z}_i}} \tag{4}$$

Eq. 4 shows the computation of probability score where $z_i^{\wedge}$ depicts the output of last dense layer and N is number of output classes. Training is performed in adversarial manner using both of these losses hence generator loss can be described as follows.

$$loss_{GAN}(z, \hat{x})^y = \sum_{n=1}^{N} -\log\left(D_0\left(G\left(z, \hat{x}\right)\right)\right) + \sum_{n=1}^{N} -\log\left(D_1\left(G\left(z, \hat{x}\right)\right)\right) \tag{5}$$

In this Eq. 5, z denotes the input noise vector, $\hat{x}$ denotes sentence encoding and N is number of data samples. Where D_0 and D_1 are the two discriminators, one with the attention layer and the other one is without the attention layer, respectively.

IV. DATASET

In each deep learning-based technique, dataset is meant to be the backbone. If there is no standard and meaningful data, then we cannot generate accurate and precise results. For text to face synthesis, currently there is no standard dataset available. In this paper, we have also contributed to the generation of dataset. Multiple publically available datasets are explored that contain face images like Celeb [19], LFW [20] etc. Moreover, we have also generated and gathered the images of Asian people to enhance the dataset. We have defined the categories in our proposed research work based on gender, age, hair, eyes, ethnicity attributes. Dataset has following categories in gender;

1) Male
2) Female

Whereas for the age information, dataset has included following information;

1) Young
2) Adult
3) Old

TABLE 3. Statistics of self-generated and collected dataset.

Gender	Age			Eyes		Hair		Emotions		Ethnicity	
	Young	Adult	Old	Black	Brown	Black	Brown	Happy	Sad	Black	White
Male	500	500	500	500	500	500	500	500	500	500	500
Female	500	500	500	500	500	500	500	500	500	500	500

FIGURE 6. Samples frames from dataset.

For hair and eyes, following colors are considered;

1) Black
2) Brown

Dataset has been annotated on to the following emotions;

1) Happy
2) Sad

In the last for ethnicity, following attributes are selected;

1) Black
2) White
3) Brown

Dataset has been prepared by manually extracting the images from the LFW [20] and Celeb [19] dataset and then carefully annotating them using above predefined categories. A team was established for data generation purpose. This team included the five interns and two full time-domain experts. This process took approximately five weeks. We have gathered and annotated 11,000 images against defined classes. Images are pre-processed before feeding them to the proposed network.

Pre-processing involves the removal of bad quality images, resizing each image to (256,256) and image enhancement. Table 3 represents the statistics of the self-generated dataset corresponding to each class with gender classes taken as reference. Some frames from our dataset have been shown in the Fig. 6. We will soon make this dataset available for the research community.

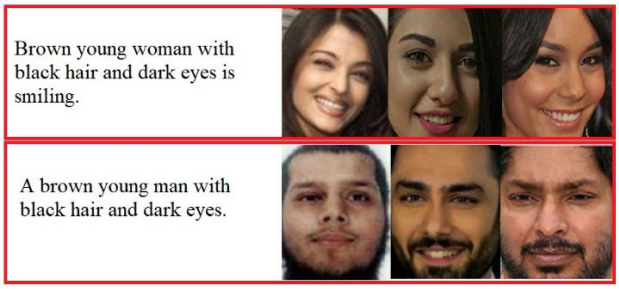


FIGURE 7. Three generated images are shown in correspondence with same input text.

TABLE 4. Comparison with other face generation model using FSD and FID Criterion.

Models	FSD	FID
AttnGAN [15]	1.269	45.56
StackGAN[4]	1.310	46.07
FTGAN [31]	1.267	44.49
Ours	1.118	42.62

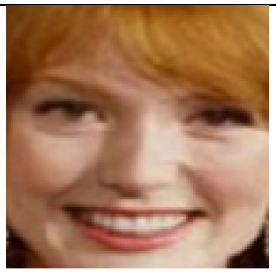
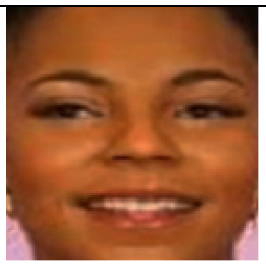
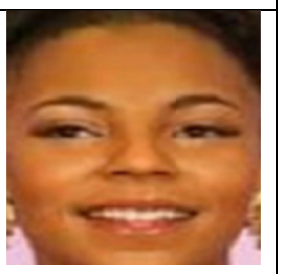
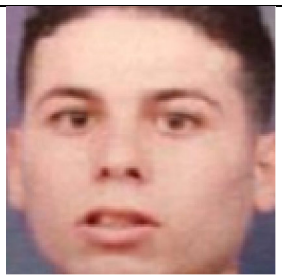
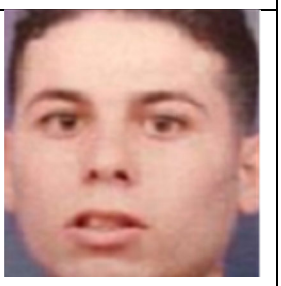
V. EXPERIMENTAL ANALYSIS

This section discusses the extensive experimental analysis that has been carried out to evaluate the performance of proposed GAN network. We have illustrated the comparison of our proposed GAN with state of the art text to face generation models and proved the efficiency of our proposed model. Later, qualitative assessment has been performed on synthesized images by human resources. The proposed network has been trained on a single Nvidia 1080Ti GPU with 11 GB memory. The model was trained for 500 epochs with initial learning rate of 0.0001. Adam optimizer is used for generator and both discriminators to optimize the weights. In the Fig.7, it is shown that our proposed model can generate photo-realistic facial images that are very near to the quality of ground truth images. Based on text, facial features of the ground truth and synthesized images are compared. Other few criterions are followed to evaluate the text to face generation. As the ultimate goal of text to face synthesis is to synthesize facial images that are correlated to ground truth images. The comparison is made by calculating the distance between the features of both images. This distance of facial features is called face semantic distance (FSD).

This distance is computed by using pre-trained FaceNet [29] F_{NET} model. FSD can be described as follows

$$FSD = \frac{1}{N}\sum_{i=1}^{N} |F_{NET}(y_i) - F_{NET}(\hat{y}_i)| \quad (6)$$

TABLE 5. Five generated images of text-to-face generation model along with the ground truth images. Left column represents the input sentences.

Sentence	Generated Face Image	Ground Truth Image
Brown young woman with black hair and dark eyes is smiling		
White middle age woman with golden hair and brown eyes is smiling		
Black young woman with black hair and dark eyes is smiling		
A brown young man with black hair and dark eyes.		
A white young man with black hair and dark eyes		

In equation 6, y_i is the generated output at each input i = 1, 2..N (N is the total number of samples) and $\hat{y}_i$ is the ground-truth image. Along with face semantic distance, we have also compared the Frechet Inception Distance (FID) [30] of synthesized images to the ground truth images. The purpose of Frechet Inception Distance is not to anticipate the

similarity of synthetic images as compared to real images. The purpose of FID score is to assess the generated images based on the statistics of group of generated images in comparison to the statistics of group of real images. FID is measured by first computing 2048 inception features from pre-trained inception v3 [36] for real and synthetic images. FID is described as

$$d^t\left(\left(m_g, C_g\right), \left(m_t, C_t\right)\right) = \left\|m_g - m_t\right\|_2^2 + Tr\left(C_g + C_t - \left(C_g C_t\right)^{\frac{1}{2}}\right) \quad (7)$$

where in equation 7, m_g and C_g are the mean and covariance from the features of generated distribution of GAN and m_t and C_t are the mean and covariance of 2048 features of ground truth sample distribution. Tr represents the trace of square matrix. Table 4 shows the comparison of proposed methodology with other GAN models. From the table 4, we interpret that the proposed model has FSD value lesser than FTGAN [31] and AttnGAN [15]. This tells us that the faces generated by our model are more similar to the ground truth face images than any of other two techniques. Moreover, the less FID score for proposed methodology tells us that mean and covariance of synthetic images by proposed model has little variation from the mean and covariance of ground truth real images. Moreover, two-staged StackGAN [4] method was also opted to show the effectiveness of proposed end-to-end trainable method. We trained StackGAN on our dataset and evaluated it based on FSD and FID measures. Other than achieving relatively higher FSD and FID values, two-staged StackGAN took longer time to converge than the proposed network. Additionally, using two discriminators proved the effect of joint adversarial learning. Facial features and entire face aesthetics are compared with real images to produce photo-realistic synthetic faces.

The synthesized images are also relevant to the textual description that is given as input. It is clear from Table 5 that images are accurately related to the description of hair, skin and eye color. To prove the efficiency of the proposed model, Table 5 is provided. The results show that the faces generated by our model are aesthetically appealing and correct regarding input sentences. Also proposed model generates images that are of higher resolution i.e. 256 × 256. There were other generated images from similar text descriptions. These samples are shown in Fig. 7.

However, Table 5 shows only those generated images that are similar to ground truth. The top results are shown in the table 5. For female face, model fuses the makeup specifications such as lip color and accessories in ground truth image that cannot be preserved. To measure the realistic high quality synthetization of proposed model, we have also monitored the

Peak Noise to Signal Ratio (PNSR). Fig. 8 shows the increase in the value of PNSR. This ratio gives the quantification of the quality of the generated image with comparison to the ground truth images. Fig. 8 shows the increase in the quality of generated images such that lower PSNR value on increasing epochs.

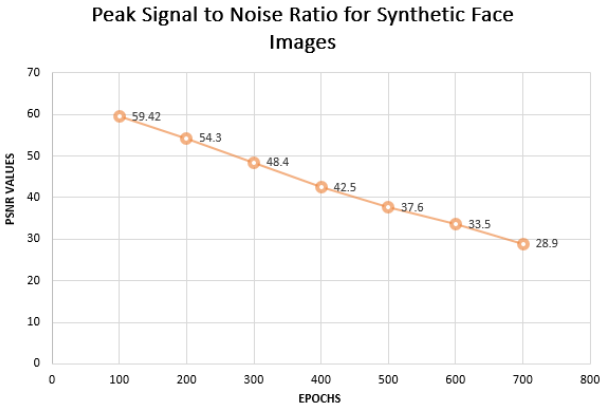


FIGURE 8. Peak signal to noise ratio over 700 epochs of the generator training.

TABLE 6. Average rating Score of Human Assessment between 1 to 5.

Volunteer	Generated Image	Generated Image and Target Description
1	4.5	4.5
2	4	4
3	5	4.5
4	4.5	5
5	5	4.5

A total of 100 synthetic facial images were shown to five human volunteers with and without textual description. Following table 6 shows the average score of the ratings they did on 100 videos. They were asked to mark the score on the scale of 1-5 based on the quality of the generated image and provided a description. It is depicted from the table 6 that the proposed architecture achieved good results based on the human assessment.

VI. CONCLUSION

In this paper, we have proposed the fully trained generative adversarial network for text to face image synthesis. The work presents a network, that trained both text encoder and image decoder for generating good quality images relative to the input sentences. By performing extensive experiments on the publicly available dataset, the superiority of our proposed methodology is proved. Moreover, in this novel task, we have also contributed towards the text to face generation dataset. Different publically available dataset along with the locally generated images have been combined. After that manual labeling of each image with defined categories has been performed. The proposed work also presents the details of the similarity between the generated faces and the ground-truth input description sentences. Experiments have shown that our proposed generative adversarial network generates natural images with good quality along with a similar face compared to the ground truth labels and faces. We compared proposed method with state of the art methods using FID and FSD scores. Proposed model achieved FSD score of 1.118 and

FID score of 42.62 that is comparatively less than other benchmark algorithms. Additionally, human ratings for our generated images are also plausible.

In future, to further improve the quality of images and to increase to similarity between the description and the generated faces, we will focus on denser and precise information related to face for the proposed architecture. This proposed work has a huge impact on security related domains like forensic analysis and public safety domain etc.

ACKNOWLEDGEMENT

The authors would also like to express our earnest gratitude to National Center of Artificial Intelligence Pakistan Fund and organization (KICS) for full supporting our research work. They also extend their gratitude to AIDA Lab CCIS Prince Sultan University Riyadh Saudi Arabia for their support to this research. The authors acknowledge support of Prince Sultan University for paying the Article Processing Charges (APC) for this publication.

REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[2] S. Hong, D. Yang, J. Choi, and H. Lee, "Inferring semantic layout for hierarchical text-to-image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7986–7994.

[3] A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, and A. Graves, "Conditional image generation with pixelcnn decoders," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 4790–4798.

[4] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "StackGAN++: Realistic image synthesis with stacked generative adversarial networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1947–1962, Aug. 2019.

[5] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," California Inst. Technol., Pasadena, CA, USA, Tech. Rep., 2011.

[6] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. 6th Indian Conf. Comput. Vis., Graph. Image Process.*, Dec. 2008, pp. 722–729.

[7] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dolláar, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.

[8] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*. [Online]. Available: http://arxiv.org/abs/1312.6114

[9] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," 2016, *arXiv:1605.05396*. [Online]. Available: http://arxiv.org/abs/1605.05396

[10] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*. [Online]. Available: http://arxiv.org/abs/1511.06434

[11] H. Dong, S. Yu, C. Wu, and Y. Guo, "Semantic image synthesis via adversarial learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5706–5714.

[12] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5907–5915.

[13] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, "Learning what and where to draw," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 217–225.

[14] S. Sharma, D. Suhubdy, V. Michalski, S. Ebrahimi Kahou, and Y. Bengio, "ChatPainter: Improving text to image generation using dialogue," 2018, *arXiv:1802.08216*. [Online]. Available: http://arxiv.org/abs/1802.08216

[15] 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," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1316–1324.

[16] T. Qiao, J. Zhang, D. Xu, and D. Tao, "MirrorGAN: Learning text-to-image generation by redescription," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1505–1514.

[17] Z. Zhang, Y. Xie, and L. Yang, "Photographic text-to-image synthesis with a hierarchically-nested adversarial network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6199–6208.

[18] A. Gatt, M. Tanti, A. Muscat, P. Paggio, R. A. Farrugia, C. Borg, K. P. Camilleri, M. Rosner, and L. van der Plas, "Face2Text: Collecting an annotated image description corpus for the generation of rich face descriptions," 2018, *arXiv:1803.03827*. [Online]. Available: http://arxiv.org/abs/1803.03827

[19] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 87–102.

[20] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," Tech. Rep., 2008.

[21] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.

[22] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*. [Online]. Available: http://arxiv.org/abs/1411.1784

[23] 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," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.

[24] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua, "Towards open-set identity preserving face synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6713–6722.

[25] 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," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2439–2448.

[26] X. Chen, L. Qing, X. He, J. Su, and Y. Peng, "From eyes to face synthesis: A new approach for human-centered smart surveillance," *IEEE Access*, vol. 6, pp. 14567–14575, 2018.

[27] X. Di and V. M. Patel, "Face synthesis from visual attributes via sketch using conditional VAEs and GANs," 2017, *arXiv:1801.00077*. [Online]. Available: http://arxiv.org/abs/1801.00077

[28] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5505–5514.

[29] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[30] 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 *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6626–6637.

[31] X. Chen, L. Qing, X. He, X. Luo, and Y. Xu, "FTGAN: A fully-trained generative adversarial networks for text to face generation," 2019, *arXiv:1904.05729*. [Online]. Available: http://arxiv.org/abs/1904.05729

[32] M. Li, W. Zuo, and D. Zhang, "Convolutional network for attribute-driven and identity-preserving human face generation," 2016, *arXiv:1608.06434*. [Online]. Available: http://arxiv.org/abs/1608.06434

[33] 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," 2017, *arXiv:1704.04086*. [Online]. Available: http://arxiv.org/abs/1704.04086

[34] T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. Tenenbaum, "Deep convolutional inverse graphics network," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2539–2547.

[35] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 4, Jul. 2017, pp. 1415–1424.

[36] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

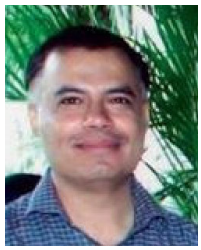
MUHAMMAD ZEESHAN KHAN received the M.S. degree in computer science from UET Lahore, Pakistan. He is currently a Team Lead with the Intelligent Criminology Lab, National Center of Artificial Intelligence, Al Khawarizmi Institute of Computer Science, UET Lahore. His areas of specialization are computer vision, machine learning, deep learning, and blockchain.

SAIRA JABEEN received the M.S. degree in computer science from UET Lahore, Pakistan. She is currently a Team Lead with the Computer Vision and Machine Learning Lab, National Center of Artificial Intelligence, Al Khawarizmi Institute of Computer Science, UET Lahore. Her area of specialization involves computer vision, machine learning, and neural networks.

MUHAMMAD USMAN GHANI KHAN received the Ph.D. degree from Sheffield University, U.K. He is currently an Associate Professor with the Department of Computer Science, University of Engineering and Technology, Lahore. He is also a Principle Investigator of the Intelligent Criminology Lab, National Center of Artificial Intelligence, Kics UET Lahore, Pakistan. His Ph.D. study was concerned with statistical modeling for machine vision signals, specifically language descriptions of video streams.

TANZILA SABA (Senior Member, IEEE) received the Ph.D. degree in document information security and management from the Faculty of Computing, Universiti Teknologi Malaysia (UTM), Malaysia, in 2012. She currently serves as a Research Professor with the College of Computer and Information Sciences, Prince Sultan University (PSU), Riyadh, Saudi Arabia. Her primary research focus in the recent years is bioinformatics, pattern recognition, machine learning, and applied soft computing. She has above two hundred publications that have around 4000 citations with H-index 40. She won best student award at the Faculty of Computing, UTM, in 2012. She received best researcher awards in PSU in 2014, 2015, 2016, and 2018. She has supervised Ph.D. and M.S. degree students. Due to her excellent research achievement, she is included in Marquis Who's Who (S & T) 2012. She is currently an editor of several reputed journals and on panel of TPC of international conferences.

ASIM REHMAT received the Ph.D. degree from the University of Engineering and Technology, Lahore, Pakistan. He is currently an Assistant Professor with the Department of Computer Science, University of Engineering and Technology, Lahore. His areas of specialization include robotics and embedded system development using artificial intelligence.



AMJAD REHMAN (Senior Member, IEEE) received the Ph.D. degree from the Faculty of Computing, Universiti Teknologi Malaysia, with a specialization in forensic documents analysis and classification, in 2010, with honor and awarded rector award for best student in the university, in 2010. He is currently a Senior Researcher with the AIDA Lab, Prince Sultan University, Riyadh, Saudi Arabia. His keen interests are in data mining, health informatics, and pattern recognition. He is author of more than 200 indexed journal articles.

USMAN TARIQ received the Ph.D. degree in information and communication technology in computer science from Ajou University, South Korea. He is a skilled Research Engineer. He has a strong background in ad hoc networks and network communications. He experienced in managing and developing projects from conception to completion. He have worked in large international scale and long-term projects with multinational organizations. He is currently with Prince Sattam bin Abdul-Aziz University as an Associate Professor with the College of Computer Engineering and Science. His research interests span networking and security fields. His current research is focused on several network security problems: botnets, denial-of-service attacks, and IP spoofing. Additionally, he is interested in methodologies for conducting security.