Generative adversarial networks with attentional multimodal for human face synthesis

Sowmya BJ[1], Meeradevi[2], Seems Shedole[3]
[1]Department of Artificial Intelligence and Data Science, Ramaiah Institute of Technology, Bangalore, India
[2]Department of Artificial Intelligence and Machine Learning, Ramaiah Institute of Technology, Bangalore, India
[3]Department of Master of Computer Applications, Ramaiah Institute of Technology, Bangalore, India

Article Info

Article history:

Received Sep 1, 2023
Revised Nov 11, 2023
Accepted Dec 5, 2023

Keywords:

Face images
Face synthesis
Generative adversarial network
High-resolution
Photorealistic

ABSTRACT

Face synthesis and editing has increased cumulative consideration by the improvement of generative adversarial networks (GANs). The proposed attentional GAN-deep attentional multimodal similarity modal (AttnGAN-DAMSM) model focus on generating high-resolution images by removing discriminator components and generating realistic images from textual description. The attention model creates the attention map on the image and automatically retrieves the features to produce various sub-areas of the image. The DAMSM delivers fine-grained image-text identical loss to generative networks. This study, first describe text phrases and the model will generate a photorealistic high-resolution image composed of features with high accuracy. Next, model will fine-tune the selected features of face images and it will be left to the control of the user. The result shows that the proposed AttnGAN-DAMSM model delivers the performance metrics like structural similarity index measure (SSIM), feature similarity index measure (FSIM) and frechet inception distance (FID) using CelebA and CUHK face sketch (CUFS) dataset. For CelebFaces attribute (CelebA) dataset, the SSIM achieves 78.82% and for CUFS dataset, the SSIM achieves 81.45% which ensures accurate face synthesis and editing compared with existing methods such as GAN, SuperstarGAN and identity-sensitive GAN (IsGAN) models.

Corresponding Author:

Sowmya BJ
Department of Artificial Intelligence and Data Science, Ramaiah Institute of Technology
Bangalore, India
Email: sowmyabj@msrit.edu

1. INTRODUCTION

The facial expression has numerous applications in several fields including healthcare, augmented reality (AR), virtual reality (VR), entertainment, and driver assistant systems [1]. The performance of existence methods depends on the quality and quantity of training information [2]. Generative adversarial networks (GANs) is a neural network architecture for generative modeling. GANs are a sort of brain network engineering that permits brain organizations to create information [3], [4]. In the past few years, they've become one of the smoking subfields in deep learning, from creating pictures of digits to photorealistic pictures of appearances [5]. GANs gain proficiency with a likelihood circulation of a dataset by setting two brain networks in opposition to one another that is Text-to-Image synthesis using GANs [6], [7]. Given a descriptive phrase of text, the model should be able to generate a photorealistic high-resolution image depicting the text up to a high accuracy [8]. Further, the model can edit selective attributes of faces individually, retaining the realistic nature of the image [9], [10]. In current years, the exciting research topic is how efficiently produce the required facial data with low costs [11]. Explore and evaluate the potential of generative adversarial networks to learn, extract and superimpose features [12]. Improving the architecture and methodology of training and usage of GANs in various tasks like extracting, classifying distinct details, and superimposing required characteristics on a given image [13], [14]. Software that uses GANs to synthesize faces from descriptive text can be used to edit the generated facial attributes according to inputs from the user [15]. Currently, there is no easy and efficient way to synthesize high-resolution face photos from text, nor is there an efficient purely computer-based way to edit facial attributes without directed inputs from the user (photoshop) [16], [17]. This paper aims to provide both of the aforementioned capabilities. Having the ability to synthesize faces and edit them efficiently with the use of the software will allow for a new age in computer-aided content creation (CaCC) and open up the world to new possibilities such as police E-Fits without requiring sketch artists. The facial expression recognition methods based on texts has limitations like the dataset utilized contains certain unwanted texts or sentences, punctuation marks that results in model computational complexity [18], [19]. Utilizing various techniques for every modality provide inconsistent classification because various nature of individual feature set utilized for every modality [20]. The facial emotions involve noise data, existence of incomplete data record and uncertainty of human emotion signals [21]. The facial expression through deep learning techniques is utilized to develop complex attention structures for extracting features [22]. But they can’t show the relationship between related position in face and expression that results in no better generalization capacity [23]. Rizkinia *et al.* [24] introduced a conditional generative adversarial network (CGAN) with color correction and total variation (TV) for Generating Indonesian Face Photos from Sketch. The developed CGAN-TV model is used in loss function for enhancing the visual quality of the output image. Then, the color correlation was employed to alter the output skin tone same to that of the ground truth. The face image dataset is composed of several resources which are corresponding to Indonesian facial features and skin tones. The developed CGAN-TV model produces better face image segmentation than an auto encoder which samples the minimum feature vector in the code layer. This dataset only utilized 10 samples for validation while the GAN requires a huge number of training data to provide better results.

Hu *et al.* [25] developed a deep learning technique that combines adversarial training and self-supervised learning for face illumination normalization (FIN-GAN). The developed FIN-GAN model is executed in dual phases. First, self-supervised learning was utilized for decay real face image to brightness and brightness-invariant module with retinex constraint. Secondly, CGAN is utilized for face image remodelling. The experiments are conducted on CAS-PEAL, Multi-PIE, YaleB, and CMU-PIE datasets. The developed FIN-GAN attains optimistic performance in face brightness normalization through numerous quantitative conditions. This developed model fails to protect personalized information. Nam *et al.* [26] implemented a low-complexity attention-generative adversarial network (LCA-GAN) for facial age evaluation with PAL and MORPH datasets. The LCA-GAN model combines a CGAN and attention architecture to the de-occlude mask-occluded face images here mouth and nose are fully sealed through mask. The developed LCA-GAN model consists of 57,118,684 parameters which indicates that it has worked an entrenched system through minimum computing sources. The mapping was complex between input to target images and provides blurred face images in mask-occluded area. Chen *et al.* [27] presented an aesthetic enhanced perception-generative adversarial network (AEP-GAN) model for Asian facial synthesis. The AEP-GAN model generates three various blocks of facial beauty such as aesthetic deformation perception, synthesis elimination and dual-agent recognition. In same period, to prevent over-beauty, the facial wedding photography dataset is used for model qualify to absorb living aesthetics. The AEP-GAN model was good with regards of an integration of quality and result of beauty in face images. The developed method was restricted by attractiveness and performance of facial rating scheme impacts certain alteration of face directly. Xia *et al.* [28] suggested a local and global perception-generative adversarial network (LGP-GAN) through dual-phase cascaded model for facial expression synthesis. The first phase employs a local network for capturing text details of critical facial areas and provides local facial areas. Then the LGP-GAN utilized a global network to absorb all facial data within second phase for providing last facial expression making locally produced output from phase one. The benefit of using this LGP-GAN is that focus on the most relevant areas and detects the interferences from other unrepresentative facial areas. The LGP-GAN method was not suitable to establish the face expression synthesis task because of several facial deformations of facial expressions.

Yan *et al.* [29] introduced an identity-sensitive generative adversarial network (IsGAN) for face photosynthesis. The cyclical-synthesized loss was developed for normalize training process and mitigate antiquity. The adversarial architecture was utilized to capture the new network loss named loss of identity recognition which was developed to preserve the complete information that was critical for photosynthesis. Additionally, to implement structural reliability during production, a cyclical-synthesized loss was employed among produced image of single domain and cyclical image of alternative. This model obtains better performance for face synthesis by using margin consequence of ArcFace. The drawback of IsGAN model requires a face identity label during training time. Wan *et al.* [30] developed generative adversarial learning for face sketch synthesis. This technique was efficient in producing entire facial contented and maintaining the face informations of input images. The higher-resolution network was changed to combine the top-level features and employed as a generator to incorporate accurate face sketch images. Additionally, style loss was assumed for detain incorporated face sketch image consume a glowing style as drawn sketch image. The benefit of using this model obtains better performance accuracy by adding facial details for portrait sketch production. The drawback of the model with incorporates sketch images have deficiency facial informations and it includes thoughtful noise properties.

Ko *et al.* [31] implemented a superstar generative adversarial network (SuperstarGAN) for image-to-image translation in large-scale domains. This SuperstarGAN model developed the ControlGAN technique of independent classifier training with data augmentation to manage overfitting problems in StarGAN classification. The generator with a trained classifier extracts small features belongings to the target domain, and SuperstarGAN obtains image-to-image translation in high-scale domains. The developed SuperstarGAN minimizes the training time because multiple models are not essential for multiple domains. The developed SuperstarGAN model performance was decreased when the high-scale domains are trained with individual models. From the overall analysis, the existing methods has limitations such as requires a huge number of training data, fails to protect personalized information, mapping was complex among input to target images and provides blurred face images. Not suitable to establish the face expression synthesis task due to the several facial deformations. Required a face identity label during training time and performance was decreased when the high-scale domains are trained with individual models. The primary scope consists of the ability to generate surrealistic human faces by generating their face with the gradual tuning of various features to closely match the description of the witness. The system would also have the capacity to understand text descriptions thereby converting it into changes in facial features for tuning processes. The primary scope of the system is to allow the synthesis of realistic human faces (images) from a given set of features (text) and to allow the user to gradually tune the features according to the user’s needs. The main aim is to allow the tuning of features to be done faster than existing systems and to control all sizes of features, be it broad like age or narrow like eye color. The final system will allow users to convert a set of text features to a face, where they will be able to use a GUI to edit features on the generated face. The primary benefits for looking at are computer-aided content creation (automating facial CGI such as de-aging) and helping police e-fits (without sketch artists, allowing users to fine-tune). The GAN contains two neural networks such as generator and discriminator which works together for face synthesis and editing tasks. The generator creates new data and the discriminator estimates how accurate those generated faces. For synthesis, the generator was trained to provide accurate faces from actual one. For editing, manipulate some features of generated face like changing hairstyle, altering age and adding glasses. The major contribution of this paper is as follows:

- The feature extractor should provide a text-to-feature function that can summarize the text into a set of features.
- The feature extractor should provide a vector function to take the set of features and generate a summary vector.
- The proposed GAN model should have a generator that can take the summary vector and output a realistic face photo.
- The two novel methods in the proposed work attentional multimodal similarity modal (AttnGAN) and deep attentional multimodal similarity model (DAMSM) are shown in the results section.

The rest of the portion present in the manuscript is organized as follows: section 2 demonstrates proposed method. Section 3 demonstrates the feature extractor-based GAN. Section 4 demonstrates the results and discussion. Section 5 demonstrates conclusion of this paper.

2. PROPOSED METHOD

Create a generative adversarial neural network to create high-resolution realistic human face photos based on the given textual features, and to allow editing of the given photos in a fast manner. The two datasets named as CelebFaces attribute (CelebA) and CUHK face sketch (CUFS) are used in this research. There are specific requirements in each sub-module which are detailed in the following sections.

2.1. CelebA dataset

The CelebA is a high-scale celebrity image face dataset which contains 200k facial images approximately, each has 40 binary features like facial expression, and hair color. In this dataset the images are utilized to cover background clutter and high pose variations. This dataset consumes rich annotations, high quantities and diversities, including:

- 10,177 identities.
- 202,599 face images.
- 5 landmark locations.
- 40 binary features per image.

This dataset has utilized as training and test sets for the computer vision tasks such as face recognition, landmark localization, attribute recognition, face synthesis and editing.

2.2. CUFS database

The CUFS dataset comprises 606 corresponding face sketch features where 188 features are form Chinese University of Hong Kong (CUHK) student dataset. 295 features are from XM2VTS dataset and 123 features from AR dataset. This paper selects 268 samples which includes 88 features from CUHK student dataset, 100 features from XM2VTS dataset and 80 features from AR dataset for training and remaining 338 features for testing purpose.

3. FEATURE EXTRACTOR BASED GENERATIVE ADVERSARIAL NETWORK

The feature extractor should provide a text-to-feature function that can summarize the text into a set of features. The feature extractor should provide a vector function to take the set of features and generate a summary vector. The proposed model focuses on generating realistic images from textual description. The focus is to generate high-resolution images by removing discriminator components. It uses text description to train the model text encoder and image decoder to generate the optimal results. The proposed AttnGAN-DAMSM has outperformed by generating high resolution images for the input textual description.

3.1. Generative adversarial network

The model should have a generator that can take the summary vector and output a realistic face photo. The model should provide a function to generate a face given any other vector that exists in its latent space. Face Recognition models have made a lot of progress in solving many real time problems in the areas of security, entertainment and many more. But the challenge is when we have a huge dataset with various poses of faces that can lead to imbalanced data. The proposed model uses GAN which focuses on multi-view faces that helps to overcome this challenge. The proposed model focuses on generating realistic images from textual description. The focus is to generate high-resolution images by removing discriminator components. The proposed work uses text description to train the model using text encoder and image decoder to generate the optimal results. The trained GAN model has outperformed by generating high resolution images for the input textual description.

3.2. Attribute editor

The editor should provide a function to tune the given feature up to the extent specified. The facial attribute editing is one of the challenging tasks for one attribute of face is predictable to be modified without affecting the other attributes. The facial attribute editing faces the issues of targeted attribute edit by manageable strength and unravelling in the presentation of attribute for reserve other attributes at the time of editing.

3.2.1. Graphical user interface

The graphical user interface (GUI) should provide a reliable, easy-to-use interface for the user to perform tasks specified by the system. The proposed model AttnGAN architecture is shown in Figure 1. The attention model creates the attention map on the image and automatically retrieves the features for generating various sub-regions of the image. The DAMSM provides fine-grained image-text identical loss for generative networks. The GAN is proposed for image synthesis from the text description [32]. Every attention architecture recovers the circumstances automatically (i.e., the most relevant word vectors) to produce various sub-areas of image; the DAMSM delivers fine-grained image-text identical loss to generative network. An AttnGAN is proposed for image synthesize among text descriptions.

The two novel methods in the proposed work AttnGAN and DAMSM are shown in the results section. They give a more significant level of command over the style of created pictures at various degrees of detail. Producing pictures from GANs isn’t new, however, more seasoned approaches depended on having the whole text input encoded as one complete vector and afterward molding the GAN on that vector. attentional generative adversarial network (AttnGAN) does likewise, yet enhances the more seasoned approaches by refining the picture in numerous stages involving word vectors too.

There are two parts proposed by the first creators of attnGAN: The first paper investigates how the creators have limited both the restrictive and unrestricted misfortunes experienced in the attentional generative organization. The DAMSM module maps words and picture districts into a typical semantic space to quantify how much the words and picture locales coordinate. The given text input is encoded with a bidirectional long short-term memory (bi-LSTM). The face synthesis and editing picture is encoded with a CNN into the text space. The process involves employing consideration of the picture encoding for every word in the sentence. Each word is associated with location-specific vectors indicating how the picture represents that particular word. Then, the DAMSM misfortune looks at the consideration vector of how the picture addresses that word to the text encoding. The entire cycle takes into consideration one thing that more seasoned approaches need, giving us fine-grained word-level data.

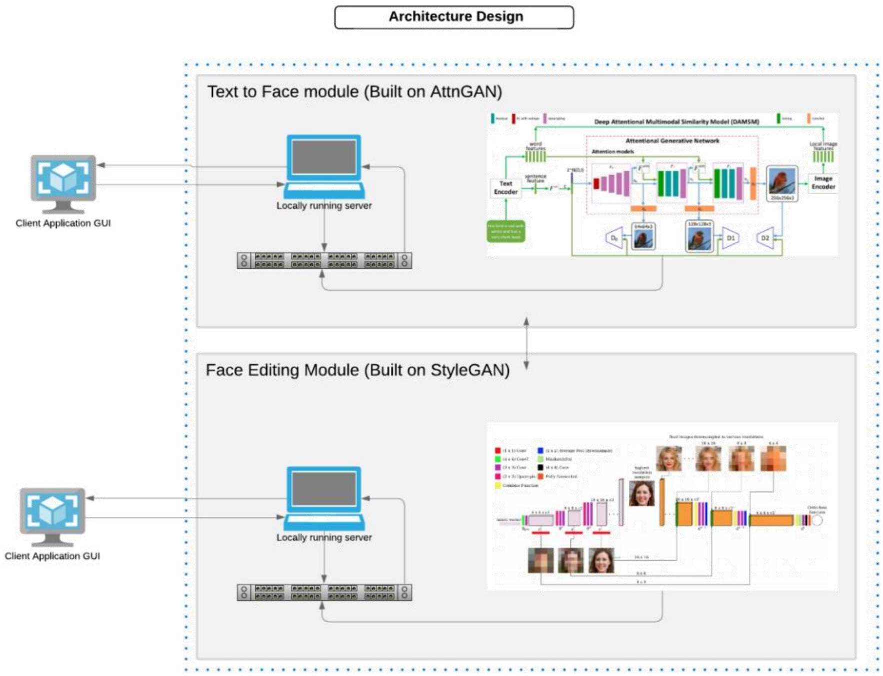

Architecture Design
Text to Face module (Built on AttnGAN)
Client Application GUI
Locally running server
Deep Attentional Multimodal Similarity Model (DAMSM)
Attentional Generative Network
Text Encoder
Image Encoder
Face Editing Module (Built on StyleGAN)
Client Application GUI
Locally running server

Figure 1. System design of the work

The text-to-face generation is built on attnGAN and trained on a subset of 20k images from the CelebA dataset. The face editing module uses a collection of datasets, particularly images with corresponding text labels, with the extent of expression. StyleGAN2 encoder is used to encode the images, and feature direction vectors are retrieved from classified latent vectors. For the feature extraction and face generation module, the user initially is given a text box to put in a few sentences resembling a description of a face. When the user clicks on the submit button, the text is first saved to a text file, and the module starts running.

Once the text has been saved, the word features are extracted to form a sentence vector. The first attention model calculates a hidden state as a function of the sentence vector and random noise. Starting from the hidden state, an initial low-resolution image is produced. Subsequently, this image undergoes progressive refinement through multiple stages. At each stage, an additional attention model comes into play, utilizing the hidden state from the previous stage and word features. This model computes a word-context vector to comprehend novel areas within the image, determining its updated hidden state, and generating a higher-resolution image as output. For example, if the model prioritizes 'smile' during refinement, the newer image will have a more expressive smile. Once the G2 (the last attention model) has given its output (256×256) image, we display it to the user. Additionally, for each stage attention maps are generated showing the top attention features, but it isn't displayed to the user but saved nonetheless. All of the code runs on Colab notebooks, and the UI is included in the notebook.

For the face editing module, a latent space Z is mapped into the given image by a generator. By manipulating Z which is an (18,512) dimension-shaped vector that can be manipulated for various features individually to a great extent. Manipulating the latent space is a matter of vector arithmetic operations like subtraction and addition. First, choose a feature and expand its representation. Then incorporate it into the provided image vector, resulting in the enhancement of the chosen feature and the extension of its expression onto the image. The user is provided with a UI to provide images and download the modified image and latent. The user can select which feature to edit and the extent of the editing. The result is also displayed to the user.

The feature extraction and face generation modules are built on Python 2.7, and use PyTorch and other libraries like python-dateutil, easydict, pandas, torchfile, nltk and scikit-image. The UI is provided in the notebook running the module by using ipywidgets library. As the face generation module used attnGAN as a base, most of the code changing and module implementation was about the training and output stages. The training process used a subset of celebA, using 20,000 images. The limited computing power had locally, as well as not being able to store the vast number of images provided in the full dataset on cloud platforms for training. Due to this, it had an impact on the final face generation module, but we have seen some promising results. Given more hardware, compute power and time to train on the full dataset (we approximate about 10-12 days of training on good hardware), we expect to see better results.

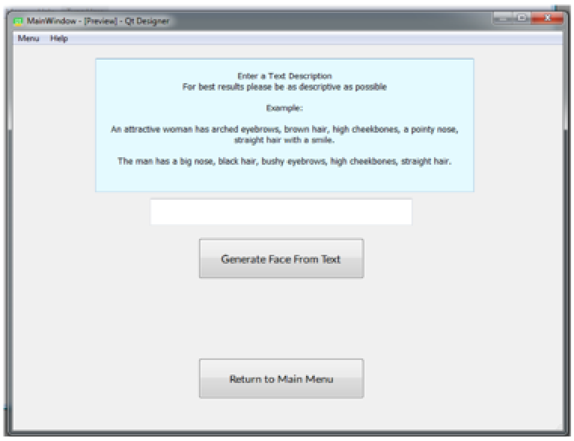
The face editing module is built on Python 3.7 by using Tensorflow. StyleGAN encoder which is built using a ResNet encoder to encode images to latent space for editing. After encoding, to extract direction vectors for each feature, image-latent vector pairs of opposite expressions of the concerned feature are classified to extract the needed normal feature direction vector. The extracted feature vectors can be used to edit facial features with regular arithmetic vector addition and subtraction. The face generation module has many problems generating valid faces due to training issues but has shown some promising results, which are detailed in later sections. The face editing module has shown very good realistic results but the tangling of features makes it uncertain of the expression of each feature for certain features.

4. RESULTS AND DISCUSSION

The user will interact with the system through the provided GUI as shown in Figures 2 and 3. All GUI text will be displayed in modern English. They have chosen to use ipywidgets (GUI widgets toolkit) to create user interface. It provides a powerful interface to create a robust GUI. Previews (mockups) are presented in Figures 2 and 3.

Start with a user who enters some text that describes a face. This description is saved into a file for processing and the module starts running once the user submits this file. The important features from the description are extracted to form a sentence vector and some noise. This is a hidden stage in the feature extraction process and from this, a low-resolution image is generated which is presented in Figures 2 and 3.

A young woman with straight brown hair wearing lipstick and arched eyebrows with narrow eyes is smiling. An example is shown in Figures 4 to 6. The Figure 4 is the initial generated image, Figure 5 is the attention map for G1 and Figure 6 is the Attention map for G2. The total time taken for this experiment is 18.107485055923462s.



Enter a Text Description
For best results please be as descriptive as possible

Example:

An attractive woman has arched eyebrows, brown hair, high cheekbones, a pointy nose, straight hair with a smile.

The man has a big nose, black hair, bushy eyebrows, high cheekbones, straight hair.

Figure 2. Initial generated

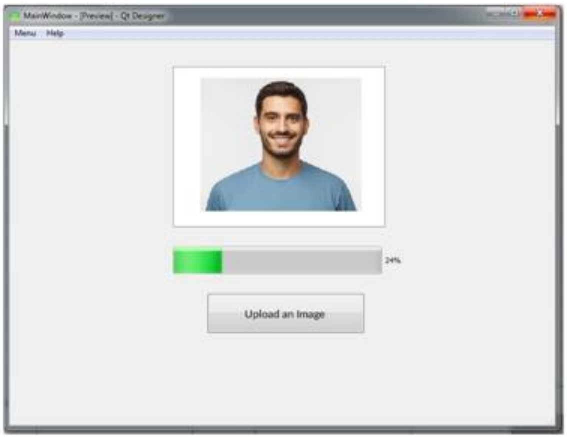


Figure 3. UI for loading the image

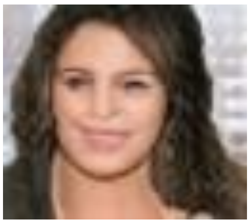


Figure 4. Initial generated image (low resolution, first hidden state, no attention): (64×64)

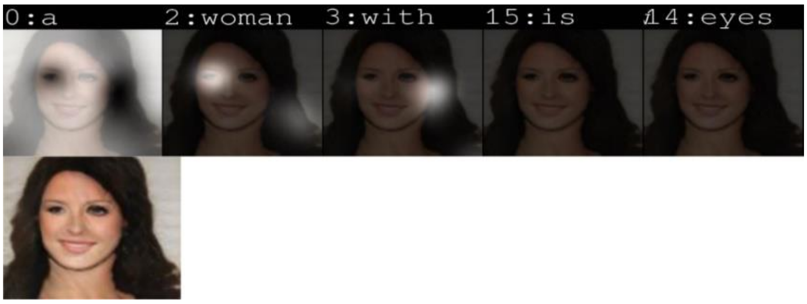


Figure 5. Attention map for G1 (first attention model) with output: (128×128)

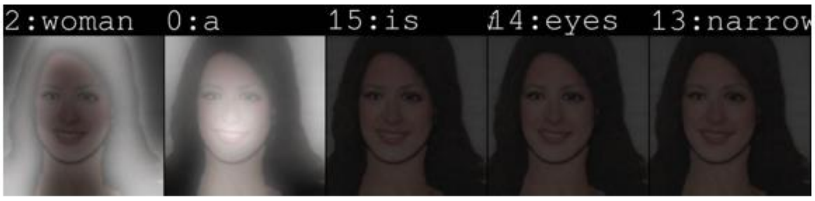


Figure 6. Attention map for G2 (second attention model) with output: (256×256)

This low-res image is then iteratively improved in stages, at each stage there is another attention model which takes the hidden state from the previous stage as well as word features and calculates a word-context vector to understand new regions in the image and calculate its new hidden state, and outputs a higher-resolution image as shown in Figures 4 to 6. The model prioritizes the word, "women" and during refinement of the image it will make the image look more feminine. The word specified on top of each image displays highest 5 most appeared words through the model for draw various sub-areas of image. The model integrates local images and correctly equivalent word-context vector for fully create the image. This method of aggregating different sub-images leads to a high-resolution image with good details.

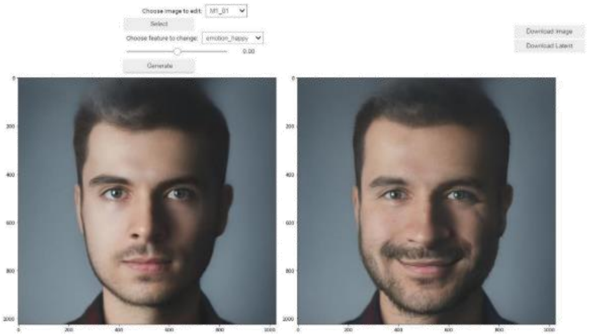
4.1. Face editing

Preprocessing and encoding for 3 very high-resolution images (>24Mb each) took about ~15 minutes on colab hardware (GPU, which is randomly assigned). The face editing module on the other hand has shown very promising results and could see used in many fields due to the time it can save in various situations. This module uses a collection of datasets, particularly images with corresponding text labels, with the extent of expression. For thi purpose, a latent space Z is mapped into the given image by a generator. By manipulating Z which is an dimension-shaped vector that can be manipulated for various features individually to a great extent.

4.2. Changing the age attribute

Afterward, the attribute governing the level of happiness on the face is modified, thereby altering the emotion of happiness. Then, using the eyes open attribute to close the eyes: (giving a negative value moves the feature the other way) as shown in Figures 7 and 8. The face editing module on the other hand has shown very promising results and could see used in many fields due to the time it can save in various situations. More research into the use of GANs for such purposes can help automate a lot of tasks to save resources in many ways. The method of finding out why the features get tangled and how to untangle them in those situations should be made a priority. A latent space Z is mapped into the given image by a generator. By manipulating Z which is a (18,512) dimension-shaped vector Then, can manipulate various features individually to a great extent. Manipulating the latent space is a matter of vector arithmetic operations like subtraction and addition. Then select the feature and its extent of expression and add it to the given image vector, to produce that feature and extent of expression on the image.

The original image is of a woman with a slight smile and open eyes. This model manipulates the image to fit the various criteria of the user. Change the features of the face such as a smile or gender, or skin colour. The properties that can be altered are the face width, face height, face shape, face age, and eyes closed or open as shown in Figures 9 and 10. This work is also used to change the emotions of the face like fear, happiness and sadness to match the description of the face to map the person more accurately.

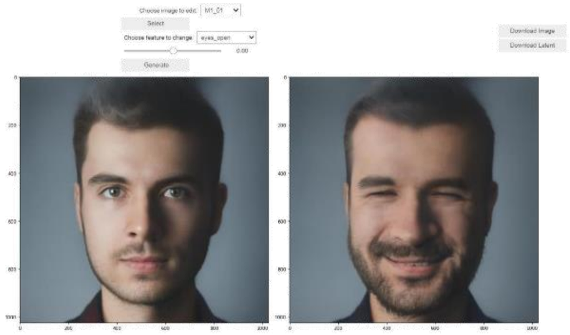


Choose image to edit: M1_01

Choose feature to change: emotion_happy

0.00

Figure 7. Face editing on happy emotion



Choose image to edit: M1_01

Choose feature to change: eyes_open

0.00

Figure 8. Face editing based on expression

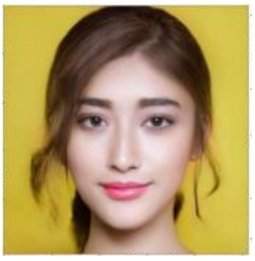


Figure 9. Original image

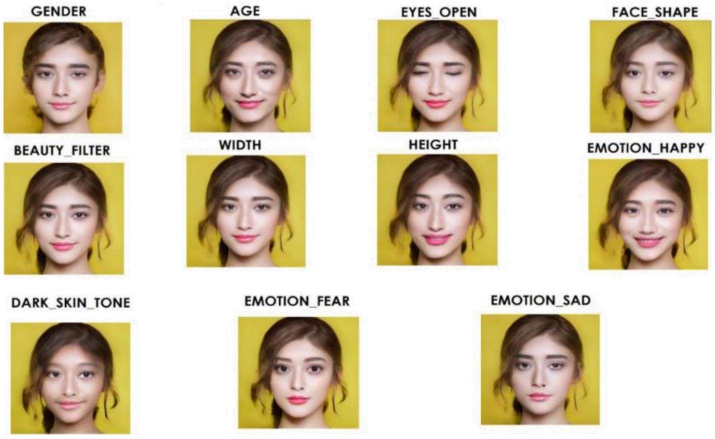


Figure 10. Attribute editing

4.3. Performance analysis

In this paper, the proposed model is stimulated using AttnGAN-DAMSM with the system requirements. The parameters like structural similarity index measure (SSIM), feature similarity index measure (FSIM) and frechet inception distance (FID) are utilized for estimating the performance of this model. The mathematical representation of these parameters is shown in (1)-(3);

SSIM:

$$SSIM = \frac{(2\mu_d\mu_o+C_1)(2\sigma_{do}+C_2)}{({\mu_d}^2+{\mu_o}^2+C_1)({\sigma_d}^2+{\sigma_o}^2+C_2)} \tag{1}$$

FSIM:

$$FSIM = \frac{\sum S_{PC}.S_G\, max(PC(M),PC(N))}{\sum max(PC(M),PC(N))} \tag{2}$$

FID:

$$FID = \left\|\mu_x - \mu_y\right\|^2 + T_x\left(\Sigma_x + \Sigma_y - 2(\Sigma_x\Sigma_y)^{1/2}\right) \tag{3}$$

Where, σ_o and μ_o denotes the standard and mean deviation of pixel score in original image respectively. σ_d and $\mu_d$denotes standard and mean deviation of pixel scores in mask-occluded image respectively. μ_{do} is the covariance of two images. C_1 and C_2 are positive constant offsets. PC is the image phase congruency.

4.3.1. Quantitative analysis

This section shows the quantitative analysis of GAN model in SSIM, FSIM and FID metrics are shown in Tables 1 and 2 respectively. Table 1 represents the quantitative analysis of various methods with CelebA dataset. Table 2 represents the quantitative analysis of various methods with CUFS dataset.

As shown in Figure 11 the performance measure of methods on CelebA dataset. The SSIM, FSIM and FID of the convolutional neural network (CNN), recurrent neural network (RNN), and gated recurrent unit (GRU) are measured and matched with this proposed AttnGAN-DAMSM model. The obtained result shows that the proposed AttnGAN-DAMSM model achieves better results by using performance metrics like SSIM, FSIM, and FID values about 78.82%, 80.63%, and 21.4 respectively while compared to other methods.

As shown in Figure 12 the performance measure of methods on CUFS dataset. The SSIM, FSIM and FID of the CNN, RNN, and GRU are measured and matched with this proposed AttnGAN-DAMSM model. The obtained result shows that the proposed AttnGAN-DAMSM model achieves better results by using performance metrics like SSIM, FSIM, and FID values about 81.45%, 83.51%, and 19.8 respectively while compared to other methods.

Table 1. Quantitative analysis of various methods for CelebA dataset

Methods	SSIM (%)	FSIM (%)	FID
CNN	73.58	75.74	29.2
RNN	75.37	77.81	26.7
GRU	76.13	79.42	23.5
AttnGAN-DAMSM	78.82	80.63	21.4

Table 2. Quantitative analysis of various methods for CUFS dataset

Methods	SSIM (%)	FSIM (%)	FID
CNN	76.69	75.49	27.4
RNN	78.58	78.85	24.7
GRU	79.32	81.64	21.9
AttnGAN-DAMSM	81.45	83.51	19.8

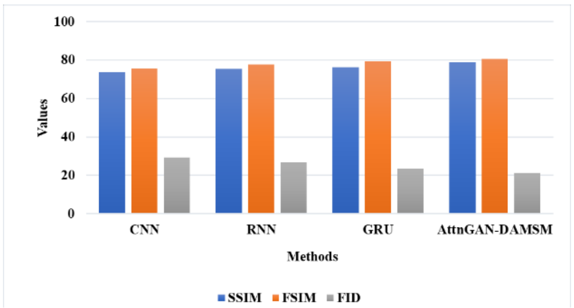


Figure 11. Performance of various methods for CelebA dataset

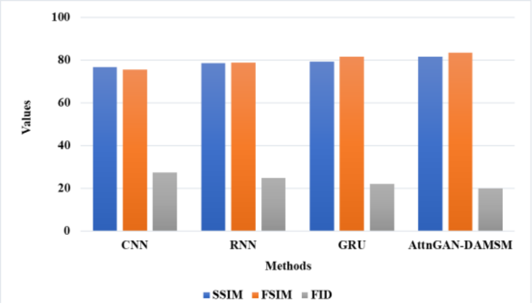


Figure 12. Performance of various methods for CUFS dataset

4.4. Comparative analysis

This section demonstrates the comparative analysis of AttnGAN-DAMSM model with performance metrics like SSIM, FSIM, and FID as shown in Tables 3 and 4. Existing research such as [29]-[31] are utilized for evaluating the ability of this model. This proposed method was trained, tested and validated using CelebA and CUFS datasets. For CelebA dataset, the SSIM was improved to 78.82%, FSIM of 80.63%. For CUFS dataset, the SSIM was improved to 81.45%, FSIM of 83.51%. The existing method has some limitations such as the IsGAN [29] model requires a face identity label during training time. The GAN [30] model incorporates drawing images have deficiency facial informations and it includes thoughtful noise properties. The SuperstarGAN [31] model performance is decreased when the high-scale domains are trained with individual models.

Table 3. Comparative analysis of proposed method with various methods for CelebA dataset

Author	Method	Dataset	Performance metrics		
			SSIM (%)	FSIM (%)	FID
Ko *et al.* [31]	Superstar GAN	CelebA	N/A	N/A	22.7
Proposed method	AttnGAN-DAMSM		78.82	80.63	21.4

Table 4. Comparative analysis of proposed method with various methods for CUFS dataset

Author	Method	Dataset	Performance metrics		
			SSIM (%)	FSIM (%)	FID
Yan *et al.* [29]	ISGAN	CUFS	75.63	77.99	N/A
Wan *et al.* [30]	GAN		N/A	73.45	N/A
Proposed method	AttnGAN-DAMSM		81.45	83.51	19.8

5. CONCLUSION

Despite the imperfect training of the face generation module, this study uncovers promising outcomes, highlighting the potential of utilizing GANs for text-to-face synthesis. Moreover, given the current advancements in the field and the availability of sufficient resources, further research holds great promise in this area. The overall time it took to train the text-to-image GAN on inferior hardware was too long to see any high-resolution results, which is just a problem of resources. The face editing module on the other hand has shown very promising results and could see used in many fields due to the time it can save in various situations. More research into the use of GANs for such purposes can help automate a lot of tasks to save resources in many ways. Methods of finding out why features get tangled, and how to untangle in those situations should be made a priority. Another key point is the absence of a proper face dataset with good labeled data. The CelebA dataset we used to have just binary labeled features, without proper captioning. The next closest would be the SCU-Face2Text, but it is sadly not a public dataset and it approximately have 400 images. The creation of a large-scale properly captioned face dataset will propagate the research and current methods forward by a long distance. The proposed AttnGAN-DAMSM delivers the performance metrics like SSIM, FSIM, and FID using CelebA and CUFS dataset. For CelebA dataset, the SSIM was improved to 78.82%, FSIM of 80.63%, and FID of 21.4. For CUFS dataset, the SSIM was improved to 81.45%, FSIM of 83.51%, and FID of 19.8. The AttnGAN-DAMSM can leads high memory and power consumption which can restrict the applicability on resource constrained devices in real-time applications. In the future, the GAN-based dimensionality reduction method is applied for improving face synthesis and face editing.

REFERENCES

[1] M. B. Shahbakhsh and H. Hassanpour, “Empowering face recognition methods using a GAN-based single image super-resolution network,” *International Journal of Engineering*, vol. 35, no. 10, pp. 1858–1866, 2022, doi: 10.5829/IJE.2022.35.10A.05.

[2] C. Yuan, K. Deng, C. Li, X. Zhang, and Y. Li, “Improving image super-resolution based on multiscale generative adversarial networks,” *Entropy*, vol. 24, no. 8, p. 1030, Jul. 2022, doi: 10.3390/e24081030.

[3] M. Sun, J. Wang, J. Liu, J. Li, T. Chen, and Z. Sun, “A unified framework for biphasic facial age translation with noisy-semantic guided generative adversarial networks,” *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 1513–1527, 2022, doi: 10.1109/TIFS.2022.3164187.

[4] X. Wu *et al.*, “F^{3}A-GAN: facial flow for face animation with generative adversarial networks,” *IEEE Transactions on Image Processing*, vol. 30, pp. 8658–8670, 2021, doi: 10.1109/TIP.2021.3112059.

[5] Y. Zeng, J. Fu, H. Chao, and B. Guo, “Aggregated contextual transformations for high-resolution image inpainting,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 7, pp. 3266–3280, Jul. 2023, doi: 10.1109/TVCG.2022.3156949.

[6] P. K. Chandaliya and N. Nain, “AW-GAN: face aging and rejuvenation using attention with wavelet GAN,” *Neural Computing and Applications*, vol. 35, no. 3, pp. 2811–2825, Jan. 2023, doi: 10.1007/s00521-022-07721-4.

[7] X. Ning, D. Gou, X. Dong, W. Tian, L. Yu, and C. Wang, “Conditional generative adversarial networks based on the principle of homologycontinuity for face aging,” *Concurrency and Computation: Practice and Experience*, vol. 34, no. 12, May 2022, doi: 10.1002/cpe.5792.

[8] X. Ma *et al.*, “An advanced chicken face detection network based on GAN and MAE,” *Animals*, vol. 12, no. 21, p. 3055, Nov. 2022, doi: 10.3390/ani12213055.

[9] Y. Liu, Q. Li, Z. Sun, and T. Tan, “A 3 GAN: an attribute-aware attentive generative adversarial network for face aging,” *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 2776–2790, 2021, doi: 10.1109/TIFS.2021.3065499.

[10] M. He, “Research on face image digital processing and recognition based on data dimensionality reduction algorithm,” *Computational Intelligence and Neuroscience*, vol. 2021, pp. 1–10, Dec. 2021, doi: 10.1155/2021/3348225.

[11] X. Li, N. Dong, J. Huang, L. Zhuo, and J. Li, “A discriminative self-attention cycle GAN for face super-resolution and recognition,” *IET Image Processing*, vol. 15, no. 11, pp. 2614–2628, Sep. 2021, doi: 10.1049/ipr2.12250.

[12] E. Pantraki and C. Kotropoulos, “Face aging using global and pyramid generative adversarial networks,” *Machine Vision and Applications*, vol. 32, no. 4, p. 82, Jul. 2021, doi: 10.1007/s00138-021-01207-4.

[13] W. Xu, X. Xie, and J. Lai, “RelightGAN: instance-level generative adversarial network for face illumination transfer,” *IEEE Transactions on Image Processing*, vol. 30, pp. 3450–3460, 2021, doi: 10.1109/TIP.2021.3061933.

[14] X. Zhang *et al.*, “Face inpainting based on GAN by facial prediction and fusion as guidance information,” *Applied Soft Computing*, vol. 111, p. 107626, Nov. 2021, doi: 10.1016/j.asoc.2021.107626.

[15] X. Zhao, W. Chen, W. Xie, and L. Shen, “Style attention based global-local aware GAN for personalized facial caricature generation,” *Frontiers in Neuroscience*, vol. 17, Mar. 2023, doi: 10.3389/fnins.2023.1136416.

[16] A. N. Razzaq, R. Ghazali, N. K. El Abbadi, and H. A. H. Al Naffakh, “Human face recognition based on local ternary pattern and singular value decomposition,” *Baghdad Science Journal*, vol. 19, no. 5, p. 1090, Oct. 2022, doi: 10.21123/bsj.2022.6145.

[17] M. Vasanthi and K. Seetharaman, “Facial image recognition for biometric authentication systems using a combination of geometrical feature points and low-level visual features,” *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 7, pp. 4109–4121, Jul. 2022, doi: 10.1016/j.jksuci.2020.11.028.

[18] M. Razzok, A. Badri, I. EL Mourabit, Y. Ruichek, and A. Sahel, “Pedestrian detection under weather conditions using conditional generative adversarial network,” *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 12, no. 4, p. 1557, Dec. 2023, doi: 10.11591/ijai.v12.i4.pp1557-1568.

[19] S. B. Sadkhan and D. J. Mardaw Zaidawi, “Geometric generative adversarial net based multiple methods for spectrum sensing in cognitive radio networks,” *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 11, no. 3, pp. 1657–1664, Jun. 2022, doi: 10.11591/eei.v11i3.3811.

[20] A. Karthik, J. Shetty, S. G., and R. Dev, “Implementation of generative adversarial networks in HPCC systems using GNN bundle,” *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 10, no. 2, p. 374, Jun. 2021, doi: 10.11591/ijai.v10.i2.pp374-381.

[21] M. A. Zaytar and C. El Amrani, “Satellite image inpainting with deep generative adversarial neural networks,” *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 10, no. 1, p. 121, Mar. 2021, doi: 10.11591/ijai.v10.i1.pp121-130.

[22] C. Kim, H. Lee, and H. Jung, “Fruit tree disease classification system using generative adversarial networks,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 3, p. 2508, Jun. 2021, doi: 10.11591/ijece.v11i3.pp2508-2515.

[23] A. M. Abood, A. R. Nasser, and H. Al-Khazraji, “Predictive maintenance of electromechanical systems based on enhanced generative adversarial neural network with convolutional neural network,” *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 12, no. 4, p. 1704, Dec. 2023, doi: 10.11591/ijai.v12.i4.pp1704-1712.

[24] M. Rizkinia, N. Faustine, and M. Okuda, “Conditional generative adversarial networks with total variation and color correction for generating indonesian face photo from sketch,” *Applied Sciences*, vol. 12, no. 19, p. 10006, Oct. 2022, doi: 10.3390/app121910006.

[25] Y. Hu, M. Lu, C. Xie, and X. Lu, “FIN-GAN: face illumination normalization via retinex-based self-supervised learning and conditional generative adversarial network,” *Neurocomputing*, vol. 456, pp. 109–125, Oct. 2021, doi: 10.1016/j.neucom.2021.05.063.

[26] S. H. Nam, Y. H. Kim, J. Choi, C. Park, and K. R. Park, “LCA-GAN: low-complexity attention-generative adversarial network for age estimation with mask-occluded facial images,” *Mathematics*, vol. 11, no. 8, p. 1926, Apr. 2023, doi: 10.3390/math11081926.

[27] H. Chen, W. Li, X. Gao, and B. Xiao, “AEP-GAN: aesthetic enhanced perception generative adversarial network for Asian facial beauty synthesis,” *Applied Intelligence*, vol. 53, no. 17, pp. 20441–20468, Sep. 2023, doi: 10.1007/s10489-023-04576-7.

[28] Y. Xia, W. Zheng, Y. Wang, H. Yu, J. Dong, and F.-Y. Wang, “Local and global perception generative adversarial network for facial expression synthesis,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1443–1452, Mar. 2022, doi: 10.1109/TCSVT.2021.3074032.

[29] L. Yan, W. Zheng, C. Gou, and F.-Y. Wang, “IsGAN: identity-sensitive generative adversarial network for face photo-sketch synthesis,” *Pattern Recognition*, vol. 119, p. 108077, Nov. 2021, doi: 10.1016/j.patcog.2021.108077.

[30] W. Wan, Y. Yang, and H. J. Lee, “Generative adversarial learning for detail-preserving face sketch synthesis,” *Neurocomputing*, vol. 438, pp. 107–121, May 2021, doi: 10.1016/j.neucom.2021.01.050.

[31] K. Ko, T. Yeom, and M. Lee, “SuperstarGAN: generative adversarial networks for image-to-image translation in large-scale domains,” *Neural Networks*, vol. 162, pp. 330–339, May 2023, doi: 10.1016/j.neunet.2023.02.042.

[32] R. Bayoumi, M. Alfonse, M. Roushdy, and A.-B. M. Salem, “Text-to-image generation based on AttnDM-GAN and DMAttn-GAN: applications and challenges,” *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 12, no. 2, pp. 1180–1188, Apr. 2023, doi: 10.11591/eei.v12i2.4199.

BIOGRAPHIES OF AUTHORS

Sowmya BJ is an Associate Professor at Ramaiah Institute of Technology’s Department of Artificial Intelligence and Data Science. Software engineering, machine learning, computer security, and data analytics are some of her interests. She can be contacted at email: sowmyabj@msrit.edu.

Meeradevi working as Associate Professor in Department of Artificial Intelligence and Machine Learning, Ramaiah Institute of Technology, Bangalore. Having 14 years of teaching and research experience. My research specialization includes networks, blockchain, artificial intelligence and machine learning. She can be contacted at email: meera_ak@msrit.edu.

Seems Shedole is Professor, Department of Master of Computer Applications, Ramaiah Institute of Technology, Bangalore, India. She obtained her Ph.D. from Visveswaraya Technological University, Belgaum. Her area of research is in Machine Learning and Bioinformatics. Her research interests include machine learning, data analytics, virtual reality and augmented reality. She is a member of ACM, and ISTE. She has published over 40 technical papers published in reputed Indian and International Conferences and Journals. She has many book chapters in her credit. She has reviewed papers of journals and the conferences. She was part of Technical Program committee and the Advisory committee of many conferences. She can be contacted at email: seema.s@msrit.edu.