Comparison of GANs for Speech to Face Generation Using a Custom Indian Face Dataset

Achyut Raghavan, Akhilesh Harkude, Udit Brahmadevara, D Ashritha, Dr. Vinodha K

Dept. of CSE, PES University, Bengaluru, Karnataka, India
(raghavanachyut@gmail.com, akhileshharkude1@gmail.com, udit.brahmadevara@gmail.com, daraashritha0703@gmail.com, vinodhak@pes.edu)

Abstract: The primary objective of this project is to develop an efficient facial generation software for forensic and police departments, addressing the critical need for accurate suspect sketching and identification. Our goal is to create user-friendly software that leverages Generative Adversarial Networks (GANs) to produce realistic facial images based on audio descriptions provided by users. Recognizing the urgency inherent in criminal investigations, our software aims to streamline the often-time-consuming process of facial sketch creation. We intend to present generated images with subtle variations and from different angles to enhance usability. By doing so, we aim to introduce a transformative tool featuring an intuitive interface and innovative functionalities. This tool has the potential to significantly improve the efficiency of criminal identification processes, thereby assisting law enforcement in solving cases more effectively and ultimately enhancing public safety.

Keywords: Audio to Text, CelebA, Criminal Detection, Deep Learning, Face Editing, GAN, MXNET, Poses of the face

1. Introduction

In the domain of criminal investigations, the accurate depiction and identification of suspects pose critical challenges for law enforcement. Conventional techniques often prove laborious and imprecise, highlighting the pressing need for a symbiotic relationship between law enforcement and artificial intelligence (AI) in today's era of rapid technological progress. This research initiative endeavors to strengthen this alliance by crafting an advanced facial sketching program explicitly tailored for law enforcement agencies. Our motivation stemmed from the glaring inefficiencies in the existing sketching system observed within a local police station. Upon investigation, it came to light that only a solitary department within the station had access to rudimentary sketching software. Furthermore, this software relied on a cumbersome drag-and-drop interface, significantly slowing down the process. To address these limitations, our methodology revolves around harnessing Generative Adversarial Networks (GANs) along with an exclusive face-refinement capability using Attn-GAN [1] by taking a textual description and generating realistic images. By leveraging these cutting-edge technologies, our primary objective is to expedite and refine the facial rendering process, thereby aiding law enforcement in swift and precise suspect identification. Ultimately, our aim is to bolster community safety by significantly enhancing the efficiency of suspect identification procedures.

The field of text-to-face generation has advanced significantly in the last few years thanks to ongoing breakthroughs in artificial intelligence. In order to close the gap between written descriptions and lifelike visual representations, researchers and developers have delved into a variety of cutting-edge technologies, such as Generative Adversarial Networks, Variational Autoencoders [2] and other advanced approaches. A variety of strategies have emerged from this broad investigation, and each has made a distinct contribution to the field's development. It is essential to acknowledge the significance of keeping up with these technical advancements to contextualize the innovative contributions that this project aims to present. Our initiative seeks to significantly improve the effectiveness of criminal identification procedures by using and expanding the capabilities of these cutting-edge technologies, contributing to the ongoing evolution of text-to-face generation and its application in law enforcement.

In our research, we aim to conduct a comprehensive comparison between two influential GANs specifically designed for distinct tasks: the Self Attention Generative Adversarial Network (SAGAN) [3] for high-quality image generation and the Deep Fusion Generative Adversarial Network (DF-GAN) [4] for synthesizing realistic images from text descriptions. With the introduction of attention-driven, long-range dependency modelling by SAGAN, detailed feature development throughout the image is made easier by enabling the generator to take into account cues from every feature location. Furthermore, SAGAN's discriminator looks for consistency in finely detailed elements in far-off areas of the picture. DF-GAN, on the other hand, eliminates entanglements between generators of different picture sizes by proposing a simpler yet more efficient text-to-image backbone. Additionally, it presents a deep text-image fusion block and a unique Target-Aware Discriminator to improve the fusion and semantic consistency of textual and visual information. By comparing SAGAN with DF-GAN, we hope to shed light on their relative contributions to the area of generative adversarial networks and identify their advantages, disadvantages, and specialized applications.

In the second and third segments of this paper, we discuss background research and related works, respectively. The fourth section goes into the many approaches we used to carry out the project. We cover how to prepare our dataset, which is an expansion of the popular CelebA-HQ [5] dataset, as well as additional preparation methods in the fifth section. The experiments conducted and the results reached are examined in the sixth and seventh sections, respectively.

2. Background

2.1 GANs:

GANs introduced in 2014 are made up of a generator and a discriminator. The objective of the discriminator is to discriminate between the real and generated data, whereas the objective of the generator is to produce realistic data, like images. In a min-max game, the role of the discriminator is to accurately classify whether the data is real or fake and the role of the generator is to fool the discriminator by creating outputs that appear authentic.

2.2 DF-GAN:

A development in the field of Generative Adversarial Networks (GANs) called Deep Fusion GAN (DF-GAN) aims to produce high-quality images by merging features from different levels of the neural network architecture. DF-GAN combines data from several network layers to improve the quality, variety, and realism of generated images. Its goal is to outperform traditional GAN models in creating visually appealing artificial content.

2.3 SA-GAN:

A GAN variation that integrates self-attention mechanisms is called SA-GAN. By enabling the network to concentrate on distinct segments of the input data, self-attention enhances its capacity to identify long-range dependencies and produce outputs that are more coherent and contextually rich.

2.4 Attn-GAN:

Another variant of GAN that highlights attention mechanisms is called Attn-GAN. Like SA-GAN, Attn-GAN makes use of attention mechanisms to enhance image generation by selectively concentrating on pertinent portions of the input while it is being generated.

2.5 SBERT:

SBERT [6] is a BERT [7] (Bidirectional Encoder Representations from Transformers) based approach for embedding and comparing sentences. It creates fixed-size sentence representations that can be utilized for a number of natural language processing tasks, including information retrieval, clustering, and sentence similarity calculations.

3. Related Work

With implications across multiple disciplines, Ian J. Goodfellow et al.'s 2014 introduction of Generative Adversarial Networks stimulated an evolutionary change in realistic image synthesis [8]. In this segment of the paper, we delve into the development of text-to-image synthesis, looking at important models that have advanced the discipline.

AttnGAN by Tao Xu et al. served as a precursor to fine-grained text-to-image generation. AttnGAN synthesised high-quality images while focusing on small textual details by applying attentional principles found in Generative Adversarial Networks.

Df-GAN introduced Deep Fusion Generative Adversarial Networks, which built upon this basis. This model creatively investigated deep fusion techniques, improving the incorporation of textual data into the image production process and leading to better synthesis results.

Zhang et al. developed Self-Attention Generative Adversarial Networks as attention mechanisms gained traction. This work showcased a higher ability to grasp long-range dependencies and improved the text-to-image synthesis process by integrating self-attention mechanisms into GAN structures.

Apart from advances in architecture, optimization techniques have been essential. Although not specifically designed for text-to-image synthesis, Kingma and Ba's Adam [9] had a major impact on training efficiency and indirectly affected the quality of the images that were produced.

The narrative goes into specialized fields like facial synthesis for humans. The authors of An Intelligent Hybrid Text-To-picture Synthesis Model, Bayoumi et al. [10], brought a distinct viewpoint to the general objective of realistic picture synthesis from textual descriptions by concentrating on producing realistic human faces.

Hou et al.'s TextFace [11] further investigated the combination of style mapping with textual descriptions for face generation and manipulation. This model offered a fresh take on text-guided image synthesis by showcasing the possibilities of mapping words to particular styles.

Zhu et al.'s DM-GAN [12] introduced the idea of memory networks for text-to-image generation in the context of dynamic memory networks. This novel method sought to improve the model's capacity to extract and apply contextual information for more complex image production.

With each model's distinct chapter, the area of text-to-image synthesis advances towards greater realism and adaptability.

4. Methodology

In this section, we outline our project methodology, which involves a comprehensive comparative analysis between DF-GAN and SA-GAN in the context of criminal face generation. In the initial segment we present the algorithm used in the process of audio to text conversion. Then we explain the GAN framework utilized for subsequent face generation based on these textual inputs. We also present two important modifications to improve the generated faces' usefulness in real-world scenarios. Using diffusion models to dynamically change the orientation of the facial features, we first study head rotation in the created faces. By addressing the need for variation in suspect face photos, this investigation adds to the dataset's realism and diversity. In the final segment, we explore how AttnGAN can be used to vary facial traits with the goal of improving the controllability and specificity of generated features that are important for criminal identification.

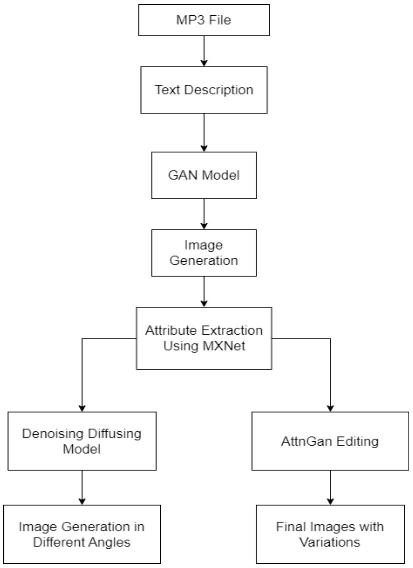

MP3 File
Text Description
GAN Model
Image Generation
Attribute Extraction Using MXNet
Denoising Diffusing Model
AttnGan Editing
Image Generation in Different Angles
Final Images with Variations

Fig 1. Workflow

4.1 Audio to Text

In our project, we recognize the significance of allowing users to submit audio files (in MP3 format) as inputs, particularly when victims or users recollect important details. This feature gives them the freedom to quickly capture important details without feeling compelled to write a description just before the procedure begins. The ability to instantly record vital facts ensures that nothing is missed or forgotten when the victim or user remembers something. This method serves scenarios when instantaneous recording is necessary for the development of an accurate facial composite by enabling the creation of a criminal sketch at any moment.

To convert audio to text, we use OpenAI's state-of-the-art Whisper ASR system [13]. This ensures accurate and contextually rich transcriptions, which serve as the basis for our GAN-based models to create detailed facial composites.

Whisper is a potent instrument that was trained on a large dataset with cutting-edge deep learning methods. It is quite accurate at transcription of spoken language; it can handle a wide range of accents, complex speech patterns, and multilingual differences found in genuine audio recordings. The accuracy and contextual richness of Whisper's transcriptions serve as the cornerstone for the next steps in our process. It guarantees the accuracy, precision, and reflection of the spoken information in the audio inputs in the textual descriptions required to create facial composites.

4.2 Caption Generation

We divided the attributes that were available in CelebA [14] into six different groups called attribute lists in order to create captions for the attributes. This classification made it easier to differentiate between similar groups of facial features and to describe them differently. Our algorithm's goal was to use a given set of attributes to produce 'N' captions. This was accomplished by choosing elements at random from attribute lists, which produced multiple variants of a given phrase. With this modification, the neural network was able to acquire knowledge of a broader range of sentences.

Every category in our algorithm is represented by a particular function that accepts matching attributes as input and outputs a meaningful sentence. There are base cases and scenarios specific to these functions. Distinct word choices are used to add variation while maintaining the integrity of the grammatical structure. Sentence structure is chosen at random for categories containing multiple attributes, and an equal probability is assigned to binary attributes.

Face Structure	**Facial Hair**	**Hairstyle**	**Facial Features**	**Appearance**	**Accessories**
Chubby	5_o_Clock_Shadow	Bald	Big_Lips	Young	Wearring_Earrings
Double_Chin	Goatee	Straight_Hair	Big_Nose	Attractive	Wearing_Hat
Oval_Face	Moustache	Wavy_Hair	Pointy_Nose	Smiling	Wearing_Lipstick
High_Cheekbones	Sideburns	Blond_Hair	Narrow_Eyes	Pale_Skin	Wearing_Necklace
		Brown_Hair	Arched_Eyebrows	Heavy_Makeup	Wearing_Necktie
		Gray_Hair	Bushy_Eyebrows	Rosy_Cheeks	Eyeglasses
		Receeding_Hairline	Mouth_Slightly_Open		

Table 1. Categories for Caption Generation

We segregate and arrange every image attribute into its appropriate category list prior to running each function. The functions use these lists as inputs, and then combine the generated descriptions into a string. The resulting string summarizes the image by combining outputs from various category functions. Through multiple iterations, this approach generates multiple captions (N captions) for the same image, offering different phrases to describe it.

The dataset is balanced by considering the number of occurrences and equalizing them in each training batch. This balancing is performed to compute the corresponding attribute weights. Finally, the summation of each attribute weight is calculated and associated to the image. Our image weights are these sums, which guarantee that each batch of attributes is represented equitably during training.

4.3 Network Architecture

For this project, we train 2 models - Self Attention (SA) GAN, Deep Fusion (DF) GAN. We referred to [15] to make specific architectural changes.

SAGAN: The SAGAN architecture in use deviates slightly from the originally mentioned design. The generator consists of two ReLU-activated, fully connected layers. The outcome of the last completely linked layer is a reduction in dimensionality of the text embeddings from 768 to 256 and finally to 100. This vector then goes through the specified layers as described in the SAGAN research after being multiplied by the input noise and reshaped to (|B|, 100, 1, 1). Before reaching the penultimate layer, the embeddings undergo processing through a fully connected layer and a ReLU-activated layer for the discriminator. Subsequently, the output vectors from preceding layers are concatenated. It is observed that the output obtained is a vector with its values between 0 and 1 when it is passed through the sigmoid layer of the convolution neural network. Further processing involves transferring this vector through two convolutional layers to generate 128 × 128-pixel images. The discriminator and generator of the SAGAN have learning rates of 0.0004 and 0.0001, respectively. The Adam optimizer [7] with $\beta 1 = 0$ and $\beta 2 = 0.9$ is used in both models.

DFGAN: The training and reference architectures of DFGAN exhibit considerable similarity, with the key difference being the generation of 128x128 images. To accommodate this reduced image size, both the generator and discriminator omit the last block responsible for handling 256x256-sized images. An enhancement to the discriminator's performance and output image quality is achieved by incorporating a matching-aware gradient policy. In terms of optimization, the DFGAN utilizes learning rates of 0.0001 for the generator and 0.0004 for the discriminator. The Adam optimizer is employed with $\beta 1 = 0$ and $\beta 2 = 0.9$.

4.4 Head Rotation

Precise manipulation of facial features without direct reference to head rotation is crucial for accurate identification in criminal face detection projects. To uphold the integrity of essential facial attributes and improve the reliability of such systems, a technique employing Denoising Diffusion Models (DDM) [16] to regulate head orientation during facial image creation was utilized.

This method generates faces by fine-tuning head orientation using Denoising Diffusion Models (DDM). First, linear regression and cluster centroids were used to compute trajectories in the latent space using a dataset (CelebA). Manipulation with the GAN generator was made possible by the classification and selection of images according to rotation angles. This generator created visible images from latent points, enabling controlled changes in head orientation. The diversity and realism of the resulting faces in different orientations while maintaining important facial features were verified by thorough analysis. Computational efficiency was optimized through the use of cluster centroids. In conclusion, this approach showed promise in the field of facial manipulation and identification by successfully addressing the challenges involved in modifying face images for controlled head orientations.

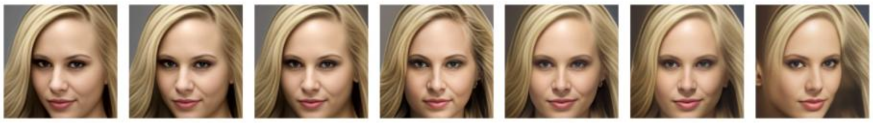


Fig 2. Multiple Angles of a Face

4.5 Attribute Variation

Towards the end of our process, we concentrate on attribute variation, adding a variety of traits to our facial composites. Using the features of the Attentional Generative Adversarial Network (AttnGAN), developed by KangHo et al. [17], we offer controlled variations in face qualities. This method allows users to precisely control and modify factors like age, gender, and expressions, making the creation of facial composites more subtle and configurable. The AttnGAN-CelebA approach offers a solid basis for achieving attribute variation as well as a simple way to increase the variety and realism of the generated faces. This tried-and-true model helps our methodology integrate more easily, improving the overall effectiveness and adaptability of our text-to-face synthesis approach.

Fig 3. Controlled Variations of Features

Original Image	Reconstruction	Bald	Bangs	Black Hair
Blonde Hair	Brown Hair	Blurry Eyebrows	Eyeglasses	Male
Mouth Slightly Open	Moustache	No Beard	Pale Skin	Young

Table 2. Controlled Variations of Features Performed for Above Figure

5. Dataset and Preprocessing

The success of machine learning models is largely dependent on the development and preparation of tailored datasets. Diverse methodologies are utilized to select datasets customized for particular study goals. This section outlines the various techniques applied in the production of datasets, including data collection, preprocessing, and labelling. One of the most popular datasets for face generation is the CelebA dataset. It has 202,599 distinct photos that are individually characterized by 40 features. The fact that this dataset lacks true diversity and that every image only belongs to celebrities is one of its drawbacks. We took on the challenge of creating our own custom dataset that resembled the CelebA dataset to cater to the needs of wanting to have Indian features for our dataset. Our dataset was combined with the CelebA-HQ subset of the CelebA dataset, which includes 30,000 of the best-labelled photos from the CelebA dataset.

5.1 Data Gathering

We collected pictures of Indians from a variety of religious and cultural backgrounds by using web scraping tools. We investigated stock photo sources featuring a diverse selection of photographs of Indian folks. We automated the process of obtaining picture urls from these websites by using the free version of Octoparse [18] web scraping application. An excel spreadsheet contained the stored URLs. These URLs were used to download the images using the tabsave Chrome extension, which downloads the image included in the URL to your local computer automatically. We were able to collect almost 30,000 raw photos as a result of this process' automation, some of which included no people and others with numerous people in a single image.

5.2 Data Preprocessing and labelling

After downloading, all of the photos were combined and put through some the preprocessing phase. We developed a pipeline that employed the Haar Cascades algorithm [19] to identify every face in a given image. We then cropped the face using OpenCV tools, resulting in a directory containing only one face per image. Then, by hand, we eliminated every picture that was blurry, had a big item in the way of the face, or lacked a face. After manual dataset cleaning, 19,045 photos remained from the original batch of images.

We experimented with a number of methods in an attempt to find the most effective way to label the processed photos. At first, we labelled 1000 photos by hand using the 40 attributes that are included in the CelebA dataset. We then fine-tuned a resnet50 [20] and a vgg16 [21] model using transfer learning techniques. To train the model on our custom-labelled data, we unfroze the final four layers of the model and fine-tuned it on our dataset, but the accuracy was not satisfactory. Finally, we labelled the 19,045 pre-processed photos using a pretrained MXNet [22] model, which achieved an accuracy of 87% on the CelebA dataset. We further had to convert the attribute file of the CelebA-HQ dataset from a text file to a csv file and then merge the images and attribute file with that of the custom dataset making sure that all of them are named in order and there is no mismatch of image to attribute.

5.3 GAN Pre processing

Before beginning the training phase, we cleaned the data using a few preprocessing techniques. To ensure that there is no disparity in the sizes of the photos, all of the images were adjusted. A few columns that didn't seem required, such as “Bags_Under_Eyes”, were removed. By giving underbalanced classes more weights, we attempted to normalize the weights for those qualities that had a significant class imbalance. Rather than training the model with the entire dataset at once, a script was added to sample a portion of the dataset and train the GAN in batches of 10%. In order to prevent any class from being overbalanced in any sample, this sampling strategy attempts to balance out the class weights.

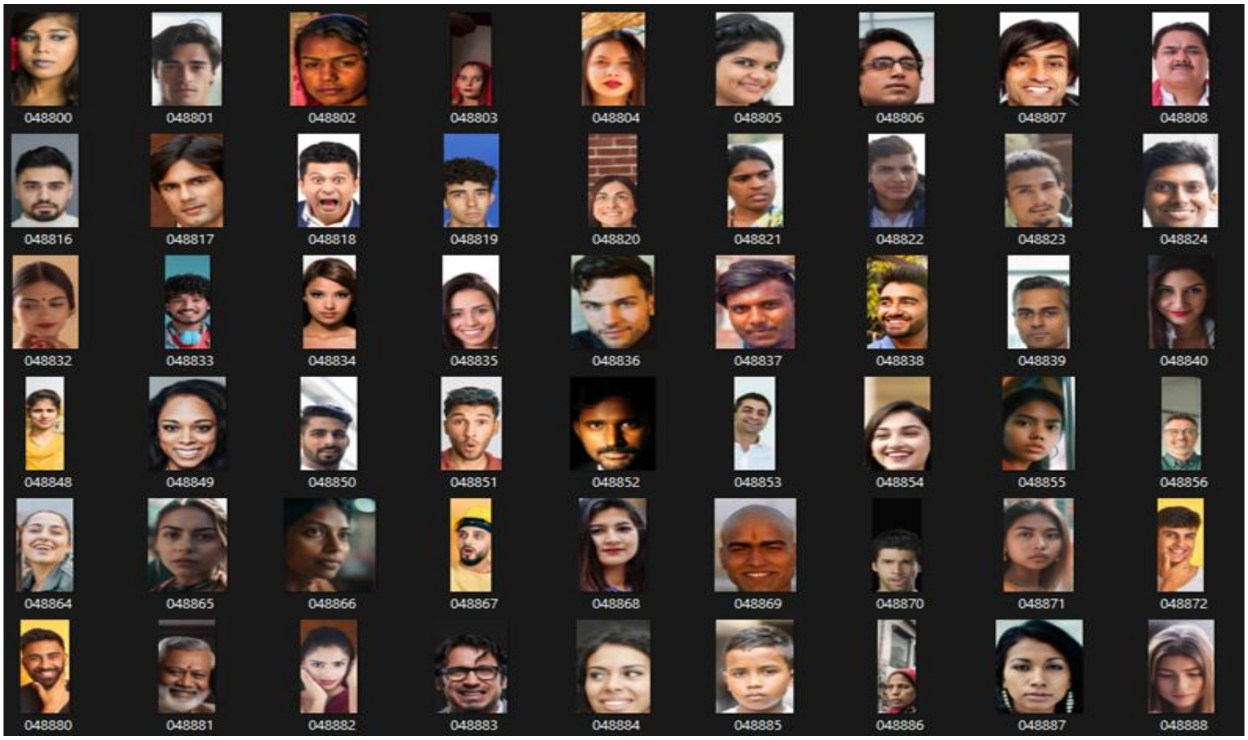
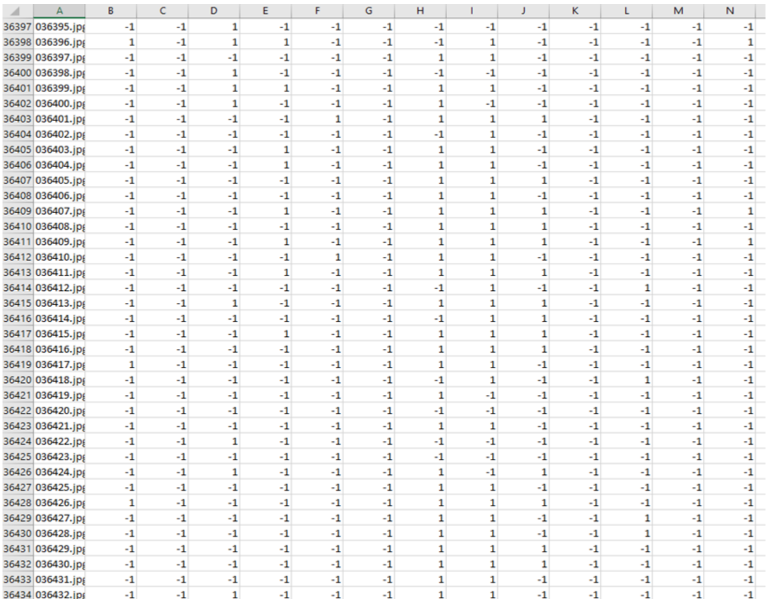


Fig 4. Cultural Infusion into CelebAHQ



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
36397	036395.jpg	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
36398	036396.jpg	1	-1	1	1	-1	-1	-1	1	-1	-1	-1	-1	1
36399	036397.jpg	-1	-1	-1	-1	-1	-1	1	1	-1	-1	-1	-1	-1
36400	036398.jpg	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
36401	036399.jpg	-1	-1	1	1	-1	-1	1	1	-1	-1	-1	-1	-1
36402	036400.jpg	-1	-1	1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1
36403	036401.jpg	-1	-1	-1	-1	1	-1	1	1	1	-1	-1	-1	-1
36404	036402.jpg	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1
36405	036403.jpg	-1	-1	-1	1	-1	-1	1	1	-1	-1	-1	-1	-1
36406	036404.jpg	-1	-1	-1	1	-1	-1	1	1	-1	-1	-1	-1	-1
36407	036405.jpg	-1	-1	-1	-1	-1	-1	1	1	1	-1	-1	-1	-1
36408	036406.jpg	-1	-1	-1	-1	-1	-1	1	1	-1	-1	-1	-1	-1
36409	036407.jpg	-1	-1	-1	1	-1	-1	1	1	1	-1	-1	-1	1
36410	036408.jpg	-1	-1	-1	-1	-1	-1	1	1	1	-1	-1	-1	-1
36411	036409.jpg	-1	-1	-1	1	-1	-1	1	1	1	-1	-1	-1	1
36412	036410.jpg	-1	-1	-1	-1	1	-1	1	1	-1	-1	-1	-1	-1
36413	036411.jpg	-1	-1	-1	1	-1	-1	1	1	1	-1	-1	-1	-1
36414	036412.jpg	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	1	-1	-1
36415	036413.jpg	-1	-1	1	-1	-1	-1	1	1	1	-1	-1	-1	-1
36416	036414.jpg	-1	-1	-1	-1	-1	-1	-1	1	1	-1	-1	-1	-1
36417	036415.jpg	-1	-1	-1	1	-1	-1	1	1	1	-1	-1	-1	-1
36418	036416.jpg	-1	-1	-1	-1	-1	-1	1	1	1	-1	-1	-1	-1
36419	036417.jpg	1	-1	-1	-1	-1	-1	1	1	-1	-1	-1	-1	-1
36420	036418.jpg	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	1	-1	-1
36421	036419.jpg	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1
36422	036420.jpg	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
36423	036421.jpg	-1	-1	-1	-1	-1	-1	1	1	-1	-1	-1	-1	-1
36424	036422.jpg	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
36425	036423.jpg	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
36426	036424.jpg	-1	-1	1	-1	-1	-1	1	-1	1	-1	-1	-1	-1
36427	036425.jpg	-1	-1	-1	-1	-1	-1	1	1	-1	-1	-1	-1	-1
36428	036426.jpg	1	-1	-1	-1	-1	-1	1	1	1	-1	-1	-1	-1
36429	036427.jpg	-1	-1	-1	-1	-1	-1	1	1	-1	-1	1	-1	-1
36430	036428.jpg	-1	-1	-1	-1	-1	-1	1	1	-1	-1	1	-1	-1
36431	036429.jpg	-1	-1	-1	-1	-1	-1	1	1	1	-1	-1	-1	-1
36432	036430.jpg	-1	-1	-1	-1	-1	-1	1	1	1	-1	-1	-1	-1
36433	036431.jpg	-1	-1	-1	-1	-1	-1	1	1	-1	-1	-1	-1	-1
36434	036432.jpg	-1	-1	1	-1	-1	-1	1	1	-1	-1	-1	-1	-1

Fig 5. Annotated Images

6. Experiments and Results

When we first started training with the CelebA dataset, we ran into problems with diversity in the images that were produced. These problems were especially noticeable in the DF GAN, even after 80 epochs of training. In order to overcome this diversity challenge, we switched to a carefully selected custom dataset designed to solve these problems.

Photos from CelebA-HQ were combined with 19,045 web-scraped and labelled images to create this unique dataset. Its particular curation was intended to increase image diversity and, as a result, enhance the range of visual qualities when our models were being trained. The quality and diversity of the images were greatly improved by this shift, which yielded in consistent image generation for every model.

During the training phase, the models SA GAN and DF GAN were trained for a long period of time-roughly 12 hours. AS a result, the SA GAN completed 130 epochs, compared to the 80 of DF GAN. Table 3 compares the results we obtained while training SA and DF GAN. We generated images after training the models for 20, 50 and 80 epochs, using a common description – "The woman has long black hair. She has arched eyebrows and a pointed nose. She is young and has heavy makeup.".

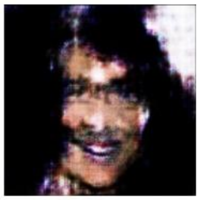
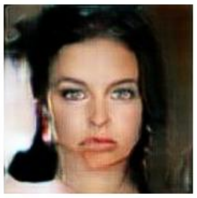
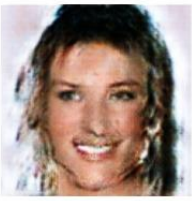
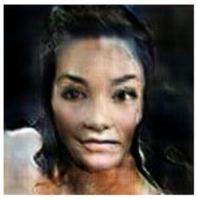
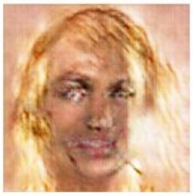
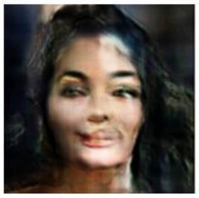
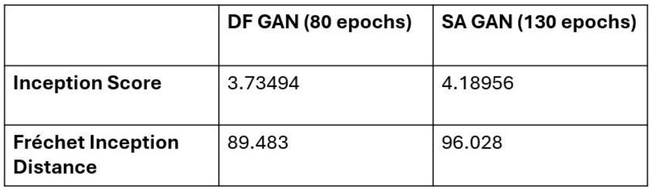
EPOCHS	SA GAN	DF GAN
20		
50		
80		

Table 3. Comparison of output of GANs

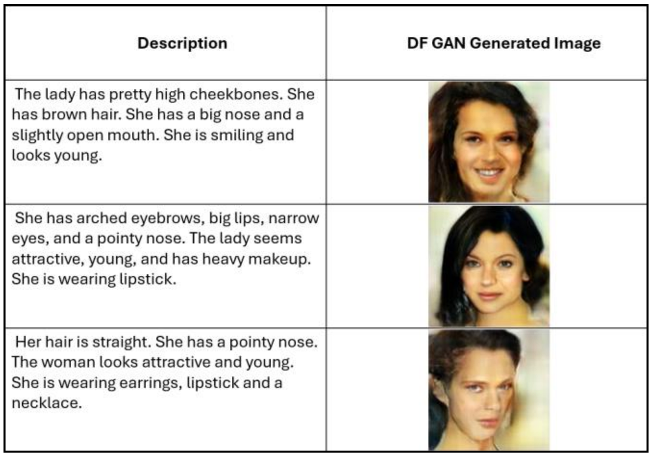
We have used the inception score [23] and Fréchet Inception Distance [24] to evaluate the accuracy of our models. The scores for obtained for these metrics are shown in the table below.



	DF GAN (80 epochs)	SA GAN (130 epochs)
Inception Score	3.73494	4.18956
Fréchet Inception Distance	89.483	96.028

Table 4. Evaluation metrics

After this, we finetuned the 70th epoch of DF GAN using the same dataset to enhance the model's performance model. We fine-tuned the dataset for another 20 epochs to achieve significantly better results.



Description	DF GAN Generated Image
The lady has pretty high cheekbones. She has brown hair. She has a big nose and a slightly open mouth. She is smiling and looks young.	
She has arched eyebrows, big lips, narrow eyes, and a pointy nose. The lady seems attractive, young, and has heavy makeup. She is wearing lipstick.	
Her hair is straight. She has a pointy nose. The woman looks attractive and young. She is wearing earrings, lipstick and a necklace.	

Table 5. DF GAN Generated Output

The resulting images strikingly mirror the provided descriptions of each individual's features. For instance, in the first row, the images display high cheekbones, a youthful appearance, and a smiling face. The second row notably captures arched eyebrows, prominent lips, narrow eyes, and a pointed nose with heavy makeup, depicting an attractive, young woman wearing lipstick. Row three showcases straight hair, a pointed nose, and an attractive, youthful appearance accessorized with earrings, lipstick, and a necklace. Overall, the generated images effectively reflect the specific facial attributes described in each case.

7. Conclusion and Future Work

An effective face sketching program has been developed as a result of the criminal departments' critical need for accuracy in suspect identification. With the use of Generative Adversarial Networks (GANs), textual descriptions are converted into facial images. A special face tuning function that makes use of Attn-GAN to improve sketch accuracy is also included. It was discovered that DF GAN performed better after extensive testing with two distinct GAN models, namely SA and DF GAN. Comparing the DF GAN to the SA GAN, the DF GAN showed superior capabilities in producing high-grade facial images with greater accuracy, better details, and finer representation of facial features.

The identification process is expedited by utilizing cutting-edge technologies and machine learning algorithms, which enhances the validity of witness testimony and solves issues faced by law enforcement organizations. This revolutionary tool seeks to greatly increase the effectiveness of criminal identification procedures through its innovative features and user-friendly interface, assisting law enforcement in solving cases more successfully and improving public safety.

The successful generation of facial images from textual descriptions holds promise in various applications, including creative content generation and facial recognition. Future research directions should prioritize the utilization of improved region-specific attributes. It is incomprehensible to have an attribute "Blonde Hair", in a nation where the majority of people have dark and black hair. Efforts ought to be focused on improving attribute representation in order to conform to distinct demographic traits, guaranteeing precision and cultural awareness. Furthermore, in order to create models that are not only technically proficient but also ethically acceptable, it is essential that computer vision specialists, cultural anthropologists, and ethicists collaborate and address biases present in training data. This comprehensive approach will aid in the creation of facial images that are precise, relevant to the context, and considerate of various cultural quirks.

References

1. Xu et al., "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 1316-1324, doi: 10.1109/CVPR.2018.00143.
2. Kingma, D. P., & Welling, M. (2019). An introduction to variational autoencoders. Foundations and Trends in Machine Learning, 12(4), 307–392. https://doi.org/10.1561/2200000056
3. Han Zhang et al., “Self-Attention Generative Adversarial Networks”, doi: arXiv:1805.08318
4. Tao, M., Tang, H., Wu, S., Sebe, N., Jing, X.Y., Wu, F., Bao, B.: “Df-gan: Deep Fusion Generative Adversarial Networks for Text-to-Image Synthesis.” arXiv preprint arXiv:2008.05865.
5. CelebA-HQ resized (256x256). (2021, April 30). Kaggle. https://www.kaggle.com/datasets/badasstechie/celebahq-resized-256x256
6. Reimers, N., & Gurevych, I.: Sentence-bert: Sentence Embeddings using Siamese Bert-networks. arXiv preprint arXiv:1908.10084 (2019)
7. Devlin, J. (2018, October 11). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv.org. https://arxiv.org/abs/1810.04805
8. Goodfellow, Ian & Pouget-Abadie, Jean & Mirza, Mehdi & Xu, Bing & Warde-Farley, David & Ozair, Sherjil & Courville, Aaron & Bengio, Y.. (2014). Generative Adversarial Networks. Advances in Neural Information Processing Systems. 3. 10.1145/3422622
9. Kingma, D. P. (2014, December 22). Adam: A method for stochastic optimization. arXiv.org. arXiv:1412.6980
10. Bayoumi, R., Alfonse, M. and Salem, A.B.M., 2021. Text-to-Image Synthesis: A Comparative Study. In Digital Transformation Technology: Proceedings of ITAF 2020 (pp. 229-251). Singapore: Springer Singapore. 10.1007/978-981-16-2275-5_14
11. Hou, X., Zhang, X., Li, Y. and Shen, L., 2022. Textface: Text-to-style mapping based face generation and manipulation. IEEE Transactions on Multimedia. 10.1109/TMM.2022.3160360
12. Minfeng Zhu, Pingbo Pan, Wei Chen, Yi Yang “DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-to-Image Synthesis”, arXiv:1904.01310 2020
13. Radford, Alec, et al. "Robust speech recognition via large-scale weak supervision." International Conference on Machine Learning. PMLR, arXiv:2212.04356 2023
14. CelebFaces Attributes (CeLEBA) Dataset. (2018, June 1). Kaggle. https://www.kaggle.com/datasets/jessicali9530/celeba-dataset
15. Deorukhkar, K., Kadamala, K., Menezes, E. (2022). FGTD: Face Generation from Textual Description. In: Ranganathan, G., Fernando, X., Shi, F. (eds) Inventive Communication and Computational Technologies. Lecture Notes in Networks and Systems, vol 311. Springer, Singapore.
16. Asperti, A. (2023, August 11). Head rotation in denoising diffusion models. arXiv:2308.06057
17. Z. {He} and W. {Zuo} and M. {Kan} and S. {Shan} and X. {Chen}, AttGAN: Facial Attribute Editing by Only Changing What You Want, 10.1109/TIP.2019.2916751
18. Web scraping tool & free web crawlers | Octoparse. (n.d.). https://www.octoparse.com/
19. Li, C., Qi, Z., Jiang, N., & Wu, J. (2017). Human face detection algorithm via Haar cascade classifier combined with three additional classifiers. IEEE. https://doi.org/10.1109/icemi.2017.8265863
20. He, K. (2015, December 10). Deep residual learning for image recognition. arXiv:1512.03385
21. Simonyan, K. (2014, September 4). Very deep convolutional networks for Large-Scale image recognition. arXiv:1409.1556
22. W. Wu. Using MXNet for Face-Related Algorithm. Accessed: May 13, 2019. [Online]. Available: https://github.com/tornadomeet/mxnet-face

23. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.:Improved Techniques for Training Gans. arXiv preprint arXiv:1606.03498 (2016)

24. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. and Hochreiter, S.: Gans Trained by a Two Time-scale Update Rule Converge to a Local Nash Equilibrium. arXiv preprint arXiv:1706.08500 (2017)