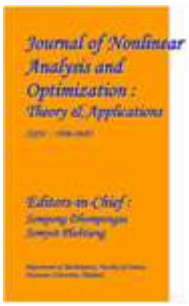
Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 15, Issue. 1, No.10 : 2024
ISSN :**1906-9685**



Synthesizing Facial Images from Text Descriptions: A Deep Fusion GAN-driven Approach

[1]L. Jagadeeswari, [1]K. Gowtham Kumar, [1]K. Mahesh, [1]G. Mahesh, [1]K. Sri Ram
[2]Dr. Mohan Mahanty
[1] UG Students, Computer science and Engineering, Vignan's Institute of Information Technology, Visakhapatnam, India.
[2]Associate Professor, Computer science and Engineering, Vignan's Institute of Information Technology, Visakhapatnam, India.

ABSTRACT: In crime investigation, the need to generate accurate facial images from textual descriptions is crucial for identifying suspects and solving cases. Existing GAN architectures, such as DCGAN (Deep Convolutional GAN) and StackGAN, have shown promise in generating realistic images. However, these models often consist of multiple generators, which can introduce ambiguity and complexity in the generation process, particularly when dealing with textual descriptions. So, we proposed a Deep Fusion Generative Adversarial Network (GAN) architecture. Our model integrates seven new layers termed as up blocks to enhance feature extraction and synthesis, while also incorporating a discriminator with Matching-sensitive gradient regularization (MS-GR) to improve the discrimination between real and generated images. Through extensive experimentation, we demonstrate the effectiveness of our methodology in producing high-quality facial images that closely align with the provided textual descriptions. For evaluation of the modal we used FID and IS as metrics. We achieved inception score of 1.318+-0.225 and FID score of 30.45.
Keywords: GAN, Text-to-Image, Deep Fusion GAN, CelebA

1.INTRODUCTION

Synthesizing realistic images from textual descriptions gives an impressive undertaking inside the realm of Deep Learning. The primary aim is to create pictures that faithfully represent the enter descriptions, executed thru the utility of Generative Adversarial Networks (GANs). While present efforts have in large part focused on generating simplistic pics like vegetation or birds from captions, this work ventures into the world of Text-to-Face technology (T2F), the subset of Text-to-Image (T2I) generation. The potential packages span various domain names which include Forensic Science, Animation, Digital Marketing, and Art.

In the landscape of artificial intelligence, GANs stand out as deep learning models, which in turn are a subset of machine learning techniques.Deep Learning is difficult in neural networks, mirrors the functioning of the human brain to figure patterns in sizable unstructured datasets. This technique differs from traditional Machine Learning algorithms, which depend upon established facts to make predictions and find styles. By harnessing the strength of Deep Learning, specially through GANs, this endeavourpursuits to push the bounds of photo synthesis, paving the manner for revolutionary packages across various industries.

The surge in popularity of Generative Adversarial Networks (GANs) stems from their exceptional ability to swiftly produce lifelike images, setting new standards for efficiency and realism in image

generation. Generative Adversarial Networks (GANs)[1] have significantly advanced the generation of highly realistic face images, facilitating their misuse in creating fake social media profiles and spreading disinformation, thus potentially causing significant repercussions. GANs offer multiple methods to create images closely resembling real ones with remarkable quality. The translation of textual descriptions into facial images holds great promise across various domains, including entertainment, virtual reality, and aiding in tasks such as locating missing persons and solving crimes.

Art encompasses a wide array of human activities and outcomes, often reflecting creative prowess and technical mastery, evoking emotions, or conveying conceptual ideas. In the process of criminal investigation, law enforcement agencies often seek assistance from artists to sketch suspects based on verbal descriptions. Anyway, employing this conventional approach can be time-intensive and may lead to significant delays in the resolution of criminal cases[2].

Leveraging Generative Adversarial Networks (GANs) for Text-to-Image (T2I) generation involves utilizing textual descriptions, particularly facial features extracted through Natural Language Processing (NLP), to produce realistic images. Natural Language Processing (NLP) is a sophisticated machine learning technology that enables computers to understand, manipulate, and comprehend human language with remarkable accuracy and precision[3].Here we can use BERT model it can act as input encoder,BERT stands as a sophisticated deep learning language model crafted to enhance the effectiveness and speed of various natural language processing (NLP) tasks[4].

1.1GENERATIVE ADVERSIAL NETWORK

GANs adopt a unique training methodology as shown in figure 1, simulating a supervised learning approach to tackle generative modelling tasks. Comprising two competing sub-models, GANs operate with a Generator and a Discriminator within their architecture. The Generator, a neural network component, is responsible for crafting data instances, while the Discriminator's role is to discern their authenticity. By evaluating whether a data instance appears genuine or fabricated, the Discriminator plays a pivotal role in refining the realism of generated outputs.

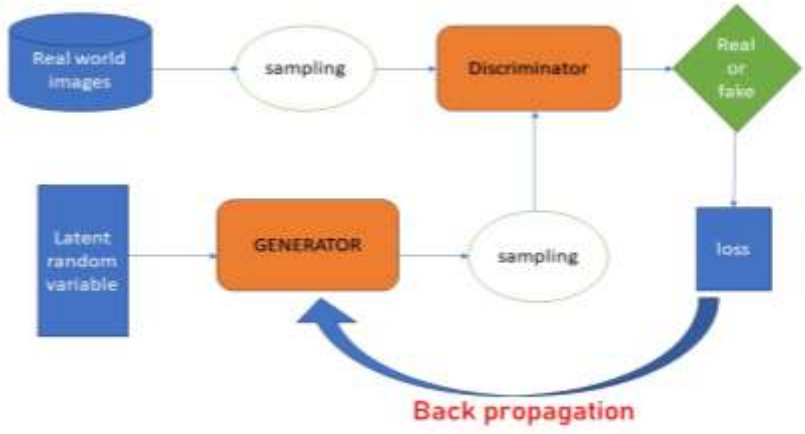


Fig 1. GAN Training Process

In the dynamics of GANs, the generator model endeavours to deceive the discriminator by continuously refining its output to appear increasingly plausible. This iterative process involves the amalgamation of backpropagation with a competitive interplay between two networks: the

Generative Network G and the Discriminative Network D. While G generates artificial images, D evaluates and categorizes them as either real or artificial, contributing to the refinement of the generative process.

The generator takes the random noise as input and try to mimic the training dataset to generate fake images and it aims to generates new samples that resembles the real data. In its role as a binary classifier, the discriminator distinguishes between genuine data samples sourced from the training dataset and counterfeit samples crafted by the generator. Simultaneously, the discriminator endeavours to accurately classify both authentic and synthetic samples, contributing to the iterative refinement process of the GAN. while the discriminator tries to correctly classify real and fake samples. If discriminator loss shows the backpropagation is, the iteration process takes place as shown in fig (1).

There are different variants of GAN namely Conditional GAN, Deep Convolutional GAN, Variational autoencoder GAN,Self-Attention GAN, Transformer GAN, Bidirectional GAN, cycle GAN, Flow-GAN and versions of style GAN.

Existing GAN models, including DCGAN, FTGAN, StyleGAN, and StackGAN, have demonstrated the ability to achieve a remarkable 57% similarity with real images[5].Here we can use Deep Fusion GAN (DFGAN), Considering the instability observed in the training processof previous FTGAN frameworks[6].

We can give input as facial Attributes, the attributes are shown in figure 2. Some of attributes are “mouth slightly open”, “smiling”, “grey hair, “No Beard”, “Bags under eyes”, “bushy eye brows”, “heavy makeup”, “oval face” etc.

In GAN technology, facial attributes such as age, gender, hair color, and facial expression are encoded as vectors and fed into the GAN model. These vectors, sourced from datasets like CelebA, guide the Generator in generating randomized facial images. This process often requires a higher number of iterations compared to earlier GAN models to produce high-quality, realistic images.

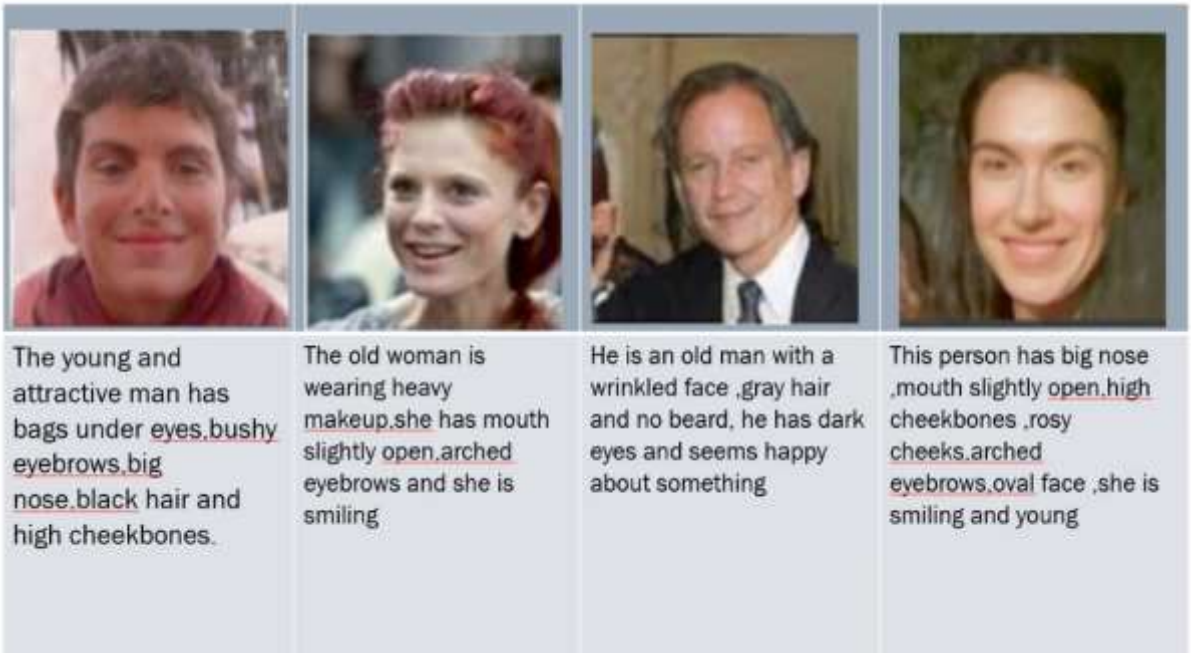


Fig 2. Features of an input image

2.LITERATURE REVIEW

Anukriti Kumar et.al and colleagues focused on generating lifelike facial images for textual descriptions through application ofsketch refinement methodology[7]. Their study highlighted the significance of utilizingCelebA dataset images, boasting a high resolution of 256×256 pixels, played a pivotal role in driving substantial improvements in the obtained results, which similarly contributed to results. They introduced the StackGAN architecture, featuring a dual-stage process, aimed at generating images with diverse facial expressions such as wide smiles or sad faces. The model exhibited promising performance, achieving a notable inception score of 4.04±0.05 across ten iterations.

Xiang Chen et.al and collaborators conducted research employing the FTGAN model[6] to synthesize realistic human faces from textual attributes describing facial features. Notably, both the image and text encoders are trained simultaneously in their approach. The FTGAN model is designed to generate images across three different scales, ranging from low resolution (e.g., 4x4 quality) to high resolution (e.g., 64x64 quality). Their model demonstrates significant enhancement in image quality, achieving a resemblance of 59% with ground truth images, and attaining an inception score of 4.61±0.05 for CUB dataset.

In their research, M. Zeeshan Khan et.al and colleagues focused on refining GANs[8] for generating realistic images by simultaneously training both image and text encoders. Their approach involved merging datasets from CelebA and LFW to enhance image quality. Notably, their model produced dual images corresponding to the same input descriptions, with a resolution of 256 x 256. The achieved quality was evidenced by FSD score of 1.218 and FID score of 44.62.

Osaid Rehman Nasir et.al and colleagues leveraged finely-grained textual descriptions to generate facial images[9]. The method they utilized encompassed the utilization of an algorithm to produce captions corresponding to images contained within the CelebA dataset. They utilized a combination of DCGAN and GANCLS loss for multimodality support. To enhance discriminator performance, they adopted a strategy of flipping labels between fake and real images, along with injecting noise. While achieving an inception score of 1.41±0.78, their method did not undergo evaluation specifically on the CelebA dataset.

Kushal Jivarajaniet.al and collaborators developed a GAN-based method for automatically synthesizing human-like faces from textual descriptions[10]. Their approach involved training a VQGAN model on the CelebA dataset while pretraining the text using a CLIP conditioning model. Notably, this model exhibited enhanced accuracy and speed in mapping text to images. However, its performance is hindered by slower processing on lower-end devices and a demand for more detailed feature descriptions in traditional images.

3. PROPOSED MODEL

In our study, we implemented the Deep Fusion GAN (DFGAN) model[11], comprises three core components: i) input Encoder, ii) Generatorand iii) Discriminator as shown in figure 3.Every component holds significant importance in the process of synthesizing images based on textual descriptions. Unlike conventional text encoders that directly convert input textual descriptions into semantic vectors, our proposed approach involves a preprocessing step. Before sentence encoding,Utilizing a well-established algorithm, we generate captions by leveraging the attributes embedded within the CelebA dataset. This distinctive preprocessing phase is designed to enrich the semantic characteristics of the produced images during the training of the model, thereby potentially elevating the accuracy and intricacy of the synthesized images.

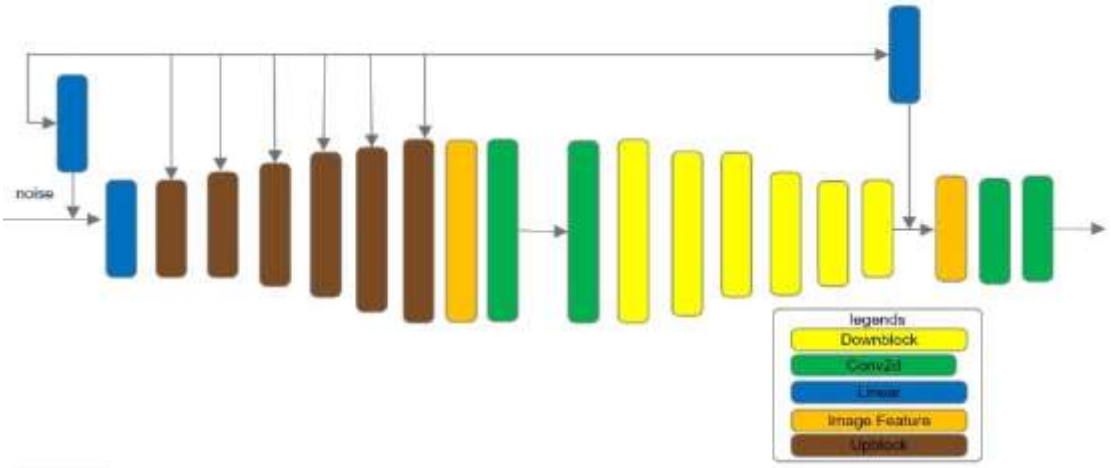


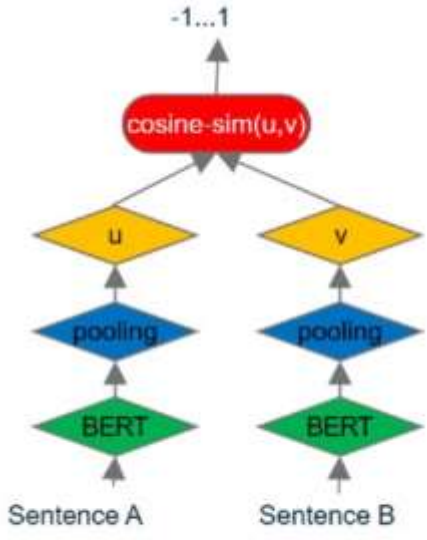
Fig 3. Deep Fusion GAN

encoder. This framework enables high-resolution image generation through a single pair of generator and discriminator, while incorporating text information and visual feature maps via. Within the DF-GAN architecture, a generator, discriminator, and multiple pre-trained text Deep Text-Image Fusion Blocks (DF Blocks) embedded within UP Blocks are introduced. Through the utilization of Matching-Aware Gradient Penalty (MA-GP) and implementation of a one-way output strategy, the model demonstrates remarkable proficiency in generating authentic images that closely align with the accompanying text descriptions.

3.1. INPUT ENCODER

In our approach, we employed Sentence BERT, a sentence-transformers model, to provide the generator with a semantic vector representing the input sentence[12]. Crafted to translate sentences and paragraphs into a 768-dimensional dense vector space, this model delivers a wide range of capabilities, spanning from clustering to semantic search, making it highly versatile in its applications. Sentence BERT, an adaptation of BERT customized for deriving semantically meaningful sentenceembeddings,facilitatesvarious NLP tasks effectively[4].

The authors showcased that conventional approaches for obtaining sentence embeddings using BERT fell short in achieving satisfactory results, particularly in tasks like textual similarity assessment. The architecture of Sentence BERT relies heavily on the training data at hand. Throughout our experimentation, we delved into diverse network structures and objective functions, aiming to enhance performance across different benchmarks.



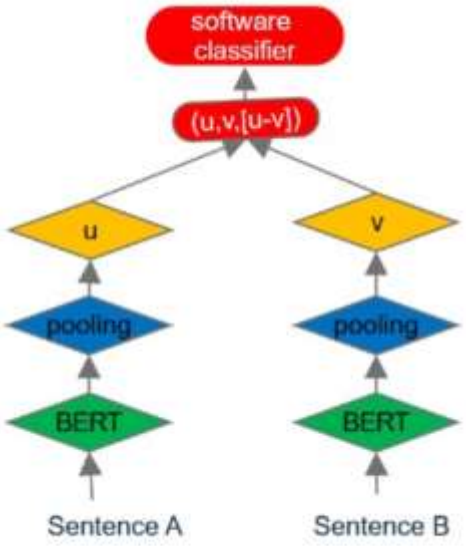


Fig 4. SBERT Regression function[12]Fig 5. SBERT Classification function[12]

Regression Function: Here cosine similarity between the embeddings of two sentences, denoted as u and v, i.e, calculated (refer to Figure 4). Mean squared error loss serves as the principal objective function guiding our computational processes.

Classification Function: We combine the sentence embeddings u and v by joining them together and including the element-wise difference |u−v|. This combined result is then multiplied by the trainable weight $W_t \in R$, which belongs to the set of real numbers 3n×k as shown figure 5

3.2. GENERATOR

Due to the inherent instability observed in GAN models, earlier text-to-image GANs often employed stacked architectures to generate high-resolution images from lower-resolution inputs.Nonetheless, the stacking of multiple generators and discriminators may introduce entanglements among different components, potentially yielding final refined images that merely resemble a blend of indistinct shapes.

Inspired by the methodology delineated insDF-GAN[13], our approach diverges from the conventional stack architecture framework. Instead, we opt for a singular generator equipped with additional layers,this facilitates the direct generating of high-resolution images from noise vector. With a focus on our generator, it works with multiple inputs: a sentence vector which is encoded by our text encoder and the noise vector sampled by Gaussian distribution, ensuring diversity in the generated images. Initially, the noise vector undergoes preprocessing via a fully connected layer and subsequent reshaping. Following this, a sequence of UP Blocks is utilized to gradually up-sample the image features. Each UP Block includes an up-sampling layer, a residual block, and DF Blocks, enabling smooth integration of text and image characteristics during the image generation procedure. Ultimately, a convolutional layer is utilized to convert the image features into concrete images.

3.2.1. UP Block

Within our DF-GAN's generator, we integrate 7 UP Blocks, each housing multiple Fusion blocks to optimize textual information utilization during fusion. Leveraging the established DeepFusion

Block (DF Block)[13], To augment fusion capabilities, we integrate a set of Affine Transformations and also RELU layers within our framework.

3.2.2. Down Block

The DF Block draws its inspiration from Conditional Batch Normalization (CBN) [14]and Adaptive Instance Normalize (ADAIN)[15]andboth of which integrate the Affine transformation[16]. However, whereas CBN and ADAIN incorporate normalization layers to align feature maps towards a normal distribution, The normalization process may contradict the goal of the Affine Transformation, which aims to amplify the distinctions among various samples.Consequently, this normalization step is omitted as it proves counterproductive for the conditional generation process.

Furthermore, the depth of our deep fusion Block enhances the text to image fusion process, enriching its capabilities.We incorporate multiple Affine layers stacked together, with a RELU layer interspersed between them. This strategy fosters the diversification of graphicalfeatures and expands the interpretation spaces, thereby accommodating a broader range of visual features corresponding to distinct text descriptions.

3.3. DISCRIMINATOR

The discriminator undertakes image processing through a sequence of Down Blocks, transforming them into image features. Afterward, the sentence vector is replicated and combined with the image features. Following this integration, an adversarial loss is calculated to assess both the realism and semantic coherence of the inputs. Through distinguishing between generated images and authentic instances, the discriminator motivates the generator to produce images of higher quality and improved semantic coherence between text and image features. Drawing inspiration from this discriminator integrates, Matching-sensitive gradient regularization(MS-GR)and One Way Output mechanism aims to steer the generator towards generating images that exhibit both heightened realism and enhanced text to image semantic coherence.

3.3.1. Matching-SensitiveGradientRegularization

In this section, we embark on a comprehensive exploration of the unconditional gradient penalty[17], offering novel insights and perspectives. Following this, we advance to further elaborate on this notion, introducing the innovative Matching-sensitive gradient regularization(MS-GR) meticulously crafted to elevate the interpretive coherence of text and images within the domain of texttoimage creation.

Based on our earlier examination, we infer that implementing gradient penalties on target data facilitates the creation of a more favourable loss landscape, thereby aiding the generator's convergence. This observation holds particular significance in the realm of textto image creation. the discriminator in texttoimage creation processes four types of inputs. To ensure semantic coherence between text and visual components, our emphasis is on applying gradient penalties to real data paired with matching text—essentially, the target data for text-to-image compilation. Consequently, within framework of MS-GR, the gradient penalty is specifically enforced on real images accompanied by matching text.

Through the integration of the MS-GR loss as regularization method within the discriminator, the model showcases enhanced convergence towards real data that harmonizes seamlessly with the provided textual context, thereby yielding generated images that closely mirror the textual descriptions. Furthermore, as the discriminator undergoes joint training within our network architecture, it effectively deters the generator from producing adversarial attributes akin to those distinct, permanent auxiliary network. Moreover, MS-GR provides the added advantage of

dispensing with the need for extra networks to ensure text to image consistency. Given that gradients are evaluated via the backpropagation procedure, the only additional computation required is the summation of gradients, rendering it computationally more efficient compared to the use of supplementary networks.

3.3.2. OneWayOutput

In prior text-to-image GANs, such as those referenced in[18][19]. one pathway determines the authenticity of the image, while the other combines the image features and vector to assess text to image semantic stability.It has come to our attention, the Two path way approach undermines the efficiency of MS-GR and impedes generator's convergence rate. Specifically, the conditional deficit generates the gradient α that points towards real and corresponding inputs following backpropagation, while the unconditional deficit yields only gradient β directed solely at real images. However, the resultant gradient, being merely the sum of α and β, fails to accurately guide towards real and corresponding data points as intended. This deviation in the final gradient, given the generator's objective of producing real and text-matching images, falls short of achieving optimal text to image semantic consistency and decelerates the generator's cohesion. Hence, we opt for the One path Way approach[13] in texttoimage analysis, as proposed in to address these shortcomings.

4. RESULTS AND ANALYSIS

This segment offers a comprehensive examination of the dataset utilized, the assessment criteria applied, and the results derived from our model.

4.1. DATASET

Our model utilizes the CelebAFaces Attributes dataset (CelebA)[20], which comprises 202,599 face images sized 178×218, featuring various celebrities. This dataset encompasses 10,177 distinct identity faces and includes 40 binary attribute annotations per image, such as arched eyebrows, attractiveness, presence of bags under the eyes, baldness, and more. Each attribute is assigned a value of 1 or -1, denoting its presence or absence in the image, respectively. Additionally, the celebA dataset provides height and width information for each image, with all images stored in JPG format.

4.2. EVALUATION METRICS

In assessing the efficiency of our network and conducting comparative analyses with existing models, we opt the evaluation metrics likeFrechet Inception Distance (FID) and alsoInception Score (IS) to analyses the performance of the propoed model with the existing state of art models.

4.2.1. FrechetInceptionDistance

The FID serves as a metric to gauge the realism and diversity of images produced by GANs. Realism refers to the extent to which generated images resemble real ones, particularly in the context of human subjects. Diversity, pertains to the degree of variation among generated images, rendering them intriguing and innovative. FID is instrumental in evaluating individual images generated by GANs, analysing the impact of modifications in neural network models on realism, and comparing the efficacy of various GAN models in image generation tasks. It effectively captures both visual quality and diversity within a single metric. A less FID score indicates a closer resemblance between generated and truth images, aiding in identifying anomalies such as additional fingers or misplaced facial features.The FID score is measuredby the following equation 1.

$$d^2 = \left|\left|mu_1 - mu_2\right|\right|^2 + Tr\left(C_1 + C_2 - 2 * \sqrt{C_1} * C_2\right) \qquad \text{Eq (1)}$$

4.2.2. Inception Score

IS definesfor widelyused metric for evaluating the images produced by GANs. This metric quantifies the realism of a GAN's output, encompassing two crucial aspects: the variety of generated outputs and the perceptual quality of each individual image. A high IS signifies that the generated images exhibit both variety and clarity, while a low score indicates deficiencies in either or both of these aspects.Therefore, a higher IS indicates the GAN's capability to produce a broad spectrum of distinct and recognizable images. As shown in equation 2 p $(y|x)$ is a conditional probability of every image.

KL = p$(y|x)$ * (log(p$(y|x)$ – log(p(y))Eq (2)

Table.1. IS and FID score calculated of different GANs

Model	**Inception score (IS)**	**FID score**
AttnGAN	1.062±0.051	41.73
StackGAN	---	46.07
DFGAN (our model)	1.318±0.225	30.45

4.3. RESULTS

The generated facial image that we have obtained through the input description as “He has a 5 o' clock shadow. His hair is black and straight. He has big lips, a big nose, bushy eyebrows and a pointy nose. The man seems attractive and young.”

The resulted image 1 as shown in figure 6.

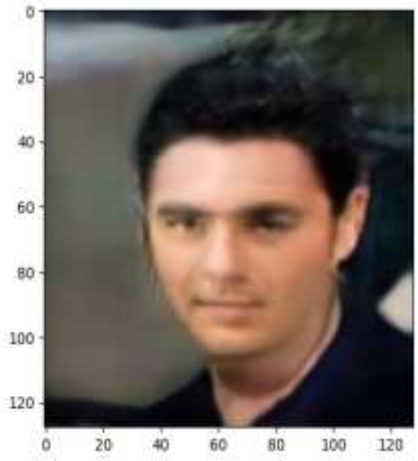


Fig 6.Generated image 1

The generated facial image that we have obtained through the input description as "The lady has pretty high cheekbones. She has brown hair. She has a big nose and aslightly open mouth. She is smiling and looks young."

The resulted image as shown in figure 7.

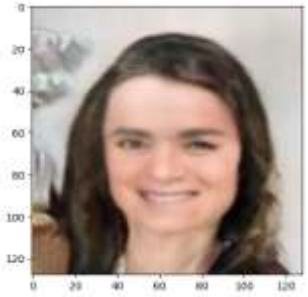


Fig 7. Generated image 2

5.CONCLUSION

Our undertaking tackles the urgent necessity of creating real looking facial representations from textual inputs, in most cases geared toward bolstering criminal investigations. Given the growing reliance on current technology in regulation enforcement, the capacity to provide particular facial images derived from textual descriptions can extensively contribute to suspect identification and crime resolution efforts.Despite improvements in textual content-to-image generation techniques, there persists a demand for heightened accuracy and realism inside the produced pictures. In response, we introduce a pioneering technique that integrates a BERT version into our GAN framework. Harnessing BERT's skillability in comprehending and reading natural language, our goal is to raise the precision and authenticity of the generated facial pictures. This goals to grant regulation enforcement businesses with greater reliable equipment for conducting criminal investigations.In our paper, we gift a pioneering DFGAN framework tailor-made for textual content-to-picture generation responsibilities. Our technique showcases a unmarried-level textual content-to-picture spine adept at directly producing high-decision pictures sans intermingling among disparate mills. Additionally, we introduce a unique Discriminator integrating Matching-sensitive gradient regularization (MS-GR) and One Way Output mechanisms, bolstering textual content-photo semantic coherence without necessitating supplementary networks. Moreover, we unveil a novel Deep Fusion Block (DF Block), facilitating greater efficient and substantial fusion of text and photograph capabilities. Extensive experimental findings validate the prevalence of our proposed DF-GAN framework over cutting-edge trendy models.Lookingin advance, improvements in GAN architectures, education methodologies, and data augmentation techniques are poised to further enrich the fidelity and realism of generated photos.

REFERENCES:

[1] X. Wang, H. Guo, S. Hu, M. C. Chang, and S. Lyu, “GAN-generated Faces Detection: A Survey and New Perspectives,” *Front. Artif. Intell. Appl.*, vol. 372, pp. 2533–2542, Feb. 2022, doi: 10.3233/FAIA230558.

[2] C. D. Frowd *et al.*, “Contemporary composite techniques: The impact of a forensically-relevant target delay,” *Leg. Criminol. Psychol.*, vol. 10, no. 1, pp. 63–81, Feb. 2005, doi: 10.1348/135532504X15358.

[3] “Natural Language Processing (Almost) from Scratch - Academic Torrents.” Accessed: Apr. 06, 2024. [Online]. Available: https://academictorrents.com/details/824fd119b03225610249c0ce6ceae778dcb7e28d

[4] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, pp. 4171–4186, Oct. 2018, Accessed: Apr. 06, 2024. [Online]. Available: https://arxiv.org/abs/1810.04805v2

[5] D. M. A. Ayanthi and S. Munasinghe, “Text-to-Face Generation with StyleGAN2,” pp. 49–64, May 2022, doi: 10.5121/csit.2022.120805.

[6] X. Chen, L. Qing, X. He, X. Luo, and Y. Xu, “FTGAN: A Fully-trained Generative Adversarial Networks for Text to Face Generation,” Apr. 2019, Accessed: Apr. 06, 2024. [Online]. Available: https://arxiv.org/abs/1904.05729v1

[7] A. Kumar, A. Mudgil, N. Dodeja, and D. K. Vishwakarma, “Realistic face generation using a textual description,” *Proc. - 5th Int. Conf. Comput. Methodol. Commun. ICCMC 2021*, pp. 917–922, Apr. 2021, doi: 10.1109/ICCMC51019.2021.9418040.

[8] M. Z. Khan *et al.*, “A Realistic Image Generation of Face from Text Description Using the

Fully Trained Generative Adversarial Networks," *IEEE Access*, vol. 9, pp. 1250–1260, 2021, doi: 10.1109/ACCESS.2020.3015656.

[9] O. R. Nasir, S. K. Jha, M. S. Grover, Y. Yu, A. Kumar, and R. R. Shah, "Text2FaceGAN: Face Generation from Fine Grained Textual Descriptions," *Proc. - 2019 IEEE 5th Int. Conf. Multimed. Big Data, BigMM 2019*, pp. 58–67, Nov. 2019, doi: 10.1109/BigMM.2019.00-42.

[10] K. Jivarajani, "Automatic Synthesis of Realistic Human Faces from Text using GANs," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 11, no. 5, pp. 7263–7271, May 2023, doi: 10.22214/IJRASET.2023.53433.

[11] M. Tao, H. Tang, F. Wu, X. Jing, B. K. Bao, and C. Xu, "DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2022-June, pp. 16494–16504, Aug. 2020, doi: 10.1109/CVPR52688.2022.01602.

[12] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *EMNLP-IJCNLP 2019 - 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf.*, pp. 3982–3992, Aug. 2019, doi: 10.18653/v1/d19-1410.

[13] M. Tao, H. Tang, F. Wu, X. Jing, B. K. Bao, and C. Xu, "DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2022-June, pp. 16494–16504, Aug. 2020, doi: 10.1109/CVPR52688.2022.01602.

[14] H. De Vries, F. Strub, J. Mary, H. Larochelle, O. Pietquin, and A. Courville, "Modulating early visual processing by language," *Adv. Neural Inf. Process. Syst.*, vol. 2017-December, pp. 6595–6605, Jul. 2017, Accessed: Apr. 08, 2024. [Online]. Available: https://arxiv.org/abs/1707.00683v3

[15] X. Huang and S. Belongie, "Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2017-October, pp. 1510–1519, Mar. 2017, doi: 10.1109/ICCV.2017.167.

[16] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 12, pp. 4217–4228, Dec. 2018, doi: 10.1109/TPAMI.2020.2970919.

[17] L. Mescheder, A. Geiger, and S. Nowozin, "Which Training Methods for GANs do actually Converge?," *35th Int. Conf. Mach. Learn. ICML 2018*, vol. 8, pp. 5589–5626, Jan. 2018, Accessed: Apr. 08, 2024. [Online]. Available: https://arxiv.org/abs/1801.04406v4

[18] T. Xu *et al.*, "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 1316–1324, Nov. 2017, doi: 10.1109/CVPR.2018.00143.

[19] H. Zhang *et al.*, "StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks," vol. 2017-Octob, pp. 5908–5916, Dec. 2016, Accessed: Apr. 08, 2024. [Online]. Available: https://arxiv.org/abs/1612.03242v2

[20] "CelebFaces Attributes (CelebA) Dataset." Accessed: Apr. 08, 2024. [Online]. Available: https://www.kaggle.com/datasets/jessicali9530/celeba-dataset