ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 08 August 2024

Text-to-Face Generation using Generative Adversarial Networks

[1] Shivani Upadhyay, [2] Dr Bharadwaja Kumar

[1] [2] School of Computer Science and Engineering (SCOPE), Vellore Institute of Technology, Chennai, India
Corresponding Author Email: [1] shivani.upadhyay2023@vitstudent.ac.in, [2] bharadwaja.kumar@vit.ac.in

Abstract— ***This project introduces a groundbreaking method in deep learning, leveraging Generative Adversarial Networks (GANs) to convert textual descriptions into high-resolution human facial images. Our approach integrates the textual and visual domains through a sophisticated model combining a text encoder and an image generator. This model is meticulously trained to ensure the generated images are not only visually appealing but also contextually precise, thereby surpassing the performance of existing methods in terms of quality, diversity, and consistency. The technical infrastructure of our project relies on a robust implementation of a Deep Convolutional Generative Adversarial Network (DCGAN) and the use of PyTorch, facilitating the complex processing required for transforming textual inputs into facial images. This pioneering work does not merely advance the state of the art in text-to-face synthesis but establishes a new paradigm for multimodal content generation. By integrating natural language understanding with visual content creation, our method paves the way for innovative applications across diverse fields such as entertainment, gaming, and assistive technologies. This represents a significant stride towards enabling more seamless interaction between humans and machines, fulfilling critical needs in areas like cosmetic surgery planning, forensic reconstruction, and the creation of personalized avatars in digital environments.***

Keywords— ***Deep Learning, Generative Adversarial Networks (GANs), Textual Descriptions, High-Resolution Human Facial Images, Text Encoder, Image Generator, Quality, Diversity, Consistency, Deep Convolutional Generative Adversarial Network (DCGAN), PyTorch, Text-to-Face Synthesis, Multimodal Content Generation, Natural Language Understanding, Visual Content Creation, Entertainment, Gaming, Assistive Technologies, Cosmetic Surgery Planning, Forensic Reconstruction, Personalized Avatars, Digital Environments.***

I. INTRODUCTION

These discussions encompass a wide array of methodologies and approaches aimed at enhancing the quality and accuracy of images generated from textual descriptions. A model called TTF-HD is introduced, which utilizes a multi-label classifier to enhance text-to-face image synthesis by closely integrating textual descriptions with facial attribute generation, showing promising advancements in the accuracy and quality of generated images [1]. A new model excels in converting text to highly realistic images through a refined encoding process, highlighting significant improvements in image quality and fidelity to original text descriptions [2]. Further development of GAN technology improves the interaction between text inputs and image outputs, enhancing the reliability and quality of generated images [3]. Advancements in generating multiple image resolutions from textual descriptions significantly aid various applications such as augmented reality and virtual scene generation [4]. Challenges and solutions in text-to-image synthesis are investigated with a focus on maintaining high fidelity between generated images and textual descriptions, leading to more precise visual representations [5].

Novel approaches using multiple discriminators improve the quality and accuracy of images generated from text, showcasing a substantial improvement over traditional single-discriminator GAN models [6]. Enhancements in GAN architectures support better learning of complex text descriptions to produce more detailed and accurate images, particularly in facial generation [7]. A new GAN framework integrates advanced text understanding capabilities to improve the coherence between generated images and their text descriptions, especially in the context of facial features [8]. Integration of text-to-image models with real-time applications demonstrates the practical utility of GANs in various fields such as digital media and automated content creation [9]. The use of GANs for generating facial images from detailed text descriptions shows potential in fields like security and digital identity verification [10].

A model leverages the learning from text-to-image GANs to simulate human learning processes, aiming to create more intuitive and adaptable image generation systems [11]. A multistage generation process for GANs allows for a more nuanced and layered approach to image creation from text, providing better control over the visual output [12]. A single-stream generator framework for GANs simplifies the generation process while maintaining high-quality output, particularly useful in resource-constrained environments [13]. A hybrid model combining multiple GAN architectures enhances the adaptability and accuracy of image synthesis from complex text descriptions [14]. A redescription model integrates text-to-image and image-to-text models to improve the feedback loop and accuracy of generated images [15].

A new approach to text-to-image synthesis using a simulation of human learning processes could revolutionize how machines understand and generate visual content from text [16]. Issues of generating consistent yet diverse images from textual descriptions by disentangling semantic information are addressed, which helps in producing more accurate and varied visual representations [18]. The generation of high-detail facial images from structured text for applications such as law enforcement uses a dataset with annotated facial attributes to improve the specificity of generated images [19]. Wasserstein StackGAN creates facial images from text utilizing a two-stage process that enhances the resolution and quality of the generated images [20]. All the previously mentioned works are different architectures of the generic model shown in Fig. 1.

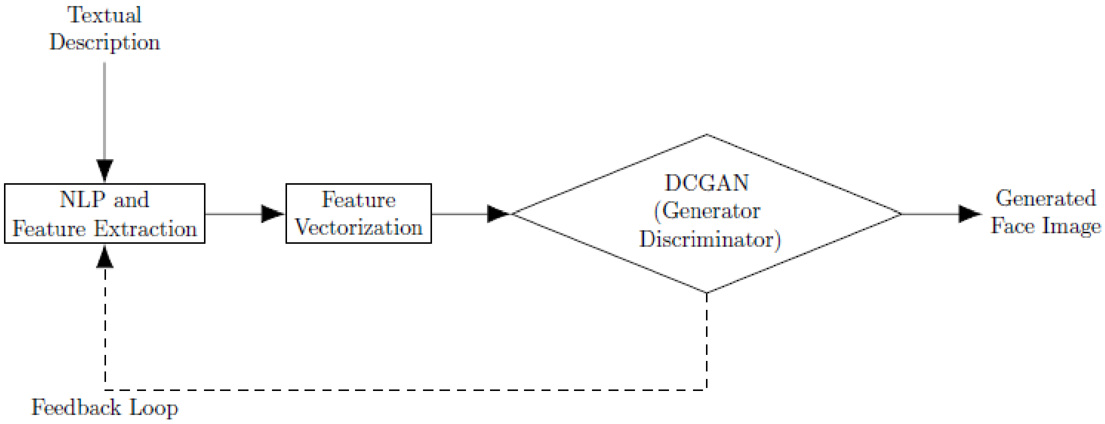


Figure 1- Generic model for generating images from text.

A Multi-Modal Attention Memory Network enhances facial attribute learning for more precise and realistic image generation from textual descriptions [21]. A method for generating high-resolution remote sensing images from textual descriptions using a structure-aware GAN improves the detail and realism of the outputs [22]. Combining VAE and GAN in a context-aware system generates high-quality images from textual descriptions, focusing on the nuances of language and image imperfections [23]. Improvements in text-to-image synthesis through knowledge-transfer techniques in the KT-GAN leverage existing image and text models to generate more coherent and contextually accurate visuals [24]. Challenges in image synthesis from text are explored, and a generative model incorporates adaptive learning techniques to improve the fidelity and detail of generated images [25]. Generating facial images from textual descriptions using StyleGAN2 focuses on the incorporation of text embeddings into the generator for more lifelike and accurate facial depictions [26].

II. METHODOLOGY

The notebook begins with the necessary imports for handling the complex operations of neural networks and image processing. Libraries such as torch for tensor operations and model building, torchvision for accessing pre-built functions and datasets for computer vision tasks, and PIL for image handling are set up. These imports ensure that all the tools required for neural network architecture, data manipulation, and image preprocessing are readily available.

Following the setup, the notebook likely progresses to prepare the dataset. In this context, the dataset would typically consist of pairs of text descriptions and corresponding facial images. These text descriptions must undergo preprocessing to be usable by the neural network. This preprocessing might involve converting the text to a numerical format through embeddings or other natural language processing techniques. Similarly, the images are preprocessed to conform to the input requirements of the neural network, which might include resizing, normalizing, and possibly augmenting the data to increase the robustness of the model. The dataset consists of pairs of textual descriptions and corresponding facial images. Text descriptions are processed through tokenization followed by embedding using a pre-trained word embedding model, transforming each text into a fixed-size vector representation. Concurrently, facial images are preprocessed to a uniform size of 64×6464×64 pixels and normalized to the range [−1,1][−1,1], in line with the activation function of the neural network.

The core of the notebook would be the construction and training of the DCGAN model, which includes a Generator and a Discriminator. The Generator's role is to create new face images from random noise input, conditioned on the text descriptions. This involves a network architecture that can take both the noise vector and the text description, process these through multiple layers (typically using transposed convolutions) and output an image that resembles those in the training dataset. The Generator G(z,t) is designed to produce an image given a noise vector ‘z’ and a text description ‘t’. The text vector is first passed through a fully connected layer to produce a text embedding, which is then concatenated with the noise vector. This combined vector is fed into a series of transposed convolutional layers that upsample the vector to generate an image. Mathematically, the generator can be represented as:

$$G(z,t) = tanh(U(z \oplus t))$$

where 'U' denotes the upsampling operations through transposed convolutions, and $\oplus$ represents the concatenation of the noise vector and text embedding.

The Discriminator, on the other hand, has the task of distinguishing between real images drawn from the dataset and fake images produced by the Generator. It assesses the authenticity of each input image and outputs a probability that the image is real. The architecture of the Discriminator typically involves convolutional layers that progressively downsample the input image, extracting features that are crucial for making this authenticity determination. The Discriminator D(x,t) aims to distinguish between real and generated images. It receives an image 'x' and a text description 't'. The image 'x' is processed through a series of convolutional layers that reduce its dimensionality, while the text description is embedded similarly to the generator's process and then spatially replicated and concatenated with intermediate convolutional features. The output of the discriminator is a scalar probability produced by a sigmoid activation function, reflecting the likelihood that 'x' is a real image conditioned on 't'. The function of the discriminator is defined as:

$$D(x,t) = \sigma(V(x \oplus t))$$

Where 'V' represents the downsampling operations through convolutional layers, and 'σ' is the sigmoid function.

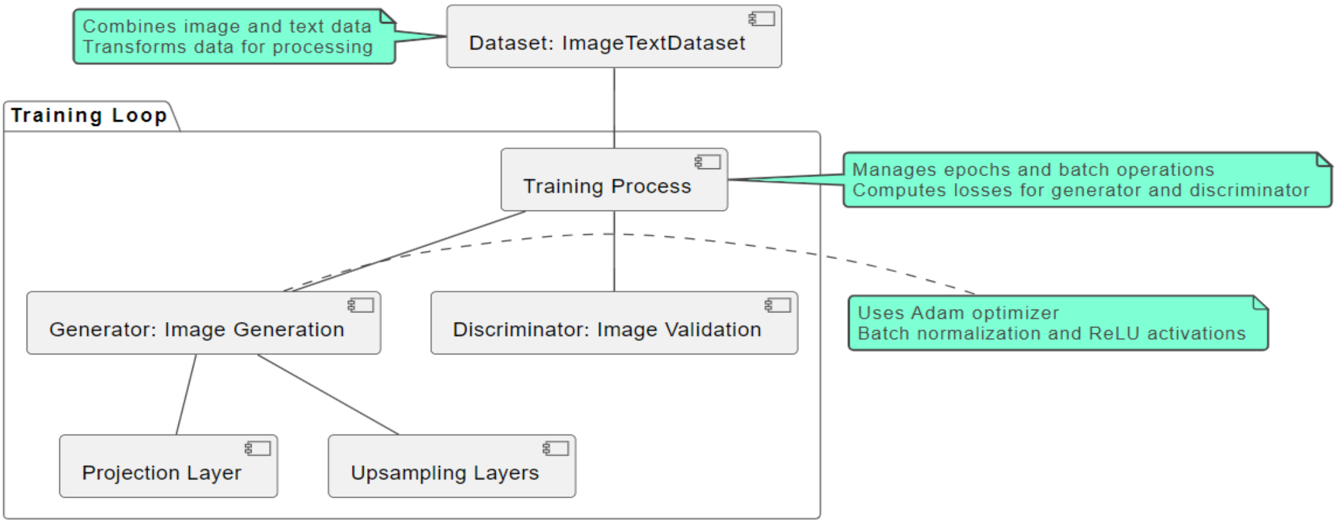

Combines image and text data
Transforms data for processing
Dataset: ImageTextDataset
Training Loop
Training Process
Manages epochs and batch operations
Computes losses for generator and discriminator
Generator: Image Generation
Discriminator: Image Validation
Uses Adam optimizer
Batch normalization and ReLU activations
Projection Layer
Upsampling Layers

Figure 2 - The architecture of the model

Training the DCGAN is a careful balance between improving the Generator and the Discriminator. The Discriminator is trained first by feeding it real images (labelled as real) and fake images from the Generator (labelled as fake). The goal here is to maximize its ability to correctly label the images. Next, the Generator is trained to fool the Discriminator by trying to produce images that are indistinguishable from real images. This adversarial training process involves alternating between training the Discriminator and the Generator with the appropriate loss functions, typically involving binary cross-entropy.

During the training process, the weights of both networks are adjusted using backpropagation based on the calculated losses. Optimizers like Adam or SGD are used to update the weights to minimize the loss, effectively improving the Generator's ability to create realistic images and the Discriminator's ability to detect fakes.

After sufficient training, the performance of the model is evaluated. This can involve qualitative assessments, where generated images are visually inspected to see how well they match the text descriptions, or quantitative assessments, using metrics designed to evaluate the quality of generated images in generative models.

The training of the DCGAN follows the standard adversarial training process where the Generator and the Discriminator are trained alternately. The objective functions for the Generator and the Discriminator are defined as follows:

- **Discriminator Objective:**

$$\left[\min_{D} E\; x \sim p[log\; D\;(x,t)] + Ez \right.$$
$$\left. \sim \mathrm{p}\left[log\left(1 - D(G(z,t),t)\right)\right]D\right]$$

- **Generator Objective:**

$$\left[\min_{\mathrm{G}} \mathrm{E}_{\mathrm{z\sim p}}\left[\log\left(1 - \mathrm{D}(\mathrm{G}(\mathrm{z},\mathrm{t}),\mathrm{t})\right)\right]\right]$$

Here, p_{data} represents the distribution of real images and text pairs, and $p_z$is the distribution of the input noise vectors. The training alternates between optimizing '*D*' to maximize the probability of assigning the correct label to both real and generated images and optimizing '*G* ' to minimize the probability of '*D*' correctly classifying the generated images as fake.

Finally, the trained model can be used to generate new images based on new text descriptions. This showcases the ability of the model to synthesize facial images from textual input, which is the culmination of the training and refining.

III. DATASET DESCRIPTION

The model uses CelebA dataset, formally known as the "Celebrity Attributes" dataset, is an extensive collection designed for the training and evaluation of algorithms dedicated to face attribute recognition. It contains over 200,000 celebrity images, meticulously annotated with 40 distinct attributes per image. These attributes encompass a wide range of facial characteristics and expressions, such as smiling, wavy hair, young, and eyeglasses, allowing for diverse and robust training in attribute detection systems.

Each image in the CelebA dataset is provided in a high-quality format, with subjects varying widely across age, ethnicity, and gender, thereby offering a comprehensive representation of human facial diversity. The images have been processed to ensure uniformity, with faces aligned and centered according to consistent landmarks (eyes, nose, and mouth). This alignment standardizes the facial features across the dataset, which is crucial for enhancing the performance of facial recognition algorithms.

Due to the extensive size of the CelebA dataset, and in the interest of managing research scope and resources, our study required the selection of a smaller, manageable subset of the total images. Out of the approximately 200,000 images available, we meticulously selected 10,000 images. This selection process was guided by the objective to maintain a uniform distribution of the annotated attributes, ensuring that each attribute is equally represented in the subset. Such a strategy allows for the analysis to remain generalizable to the larger dataset, despite the reduced number of images.

Furthermore, the subset includes the same 40 attributes as the full dataset. This consistency ensures that the integrity and breadth of the attribute annotations are preserved, providing a solid basis for our analyses. The reduction in dataset size was strategically implemented to address time constraints while still capturing a broad spectrum of facial attributes, making this subset particularly suited for focused studies aiming to explore specific aspects of facial attribute recognition.

IV. RESULTS

In a recent study, a Deep Convolutional Generative Adversarial Network (DCGAN) was employed to generate facial images from textual descriptions using the CelebA dataset, achieving notable success in text-to-image synthesis. The model produced nine distinct images, and their quality was evaluated using an Inception Score of 1.029246, indicating both novelty and fidelity. Initially, the images began as basic, abstract representations, but as training progressed, they evolved to exhibit clearer and more accurate facial features that aligned closely with the given textual descriptions.

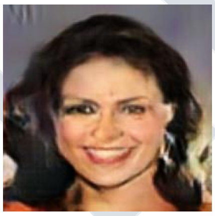
	The female has pretty high cheekbones and an oval face. She has brown hair. She has arched eyebrows and a pointy nose. She is smiling, seems attractive, young, has rosy cheeks and heavy makeup. She is wearing earrings and lipstick.
	His hair is brown and straight. He has a slightly open mouth and a pointy nose. He looks attractive and young is smiling. He is wearing a necktie.
	The male has a chubby face. He has sideburns. He has brown and straight hair. He has a big nose. The gentleman is smiling. He is wearing a necktie.
	The female has pretty high cheekbones and an oval face. Her hair is black. She has a slightly open mouth and a pointy nose. The female is smiling, looks attractive and has heavy makeup. She is wearing earrings and lipstick.

Figure 2 - sample results of our model trained on faces dataset.

The final images displayed a wide array of attributes such as different hairstyles and expressions, reflecting the model's ability to handle a diverse range of textual inputs. This diversity is particularly important for applications in fields like digital entertainment and advertising, where the ability to generate varied visual content from text can significantly enhance user engagement and content relevancy. Furthermore, the improved realism of the images by the end of training suggests that the model could generate visuals that were not only diverse but also highly detailed and comparable to real human portraits.

These results demonstrate the model's advanced capabilities in bridging natural language processing with image generation technologies, offering promising prospects for creative media and personalized content creation. This study highlights the potential of generative models to produce highly specific and varied visual content from textual descriptions, pushing the boundaries of AI-driven creative technology.

$$IS = \exp(\mathrm{E}x[DKL(p(y \mid x) \parallel p(y))])$$

Ex denotes the expectation over all generated images x. $p(y|x)$ is the conditional probability distribution of class labels y given an image x, which is estimated by the classifier.

$p(y)$ is the marginal probability distribution of the class labels over all generated images, calculated as $p(y)=\int p(y|x)dP(x)$. *DKL* denotes the Kullback-Leibler divergence, which measures how one probability distribution diverges from a second, expected probability distribution.

V. CONCLUSION

In this study, a conditional Generative Adversarial Network (cGAN) was employed to explore the capacity of generative models to produce realistic and contextually appropriate images based on conditional inputs. The framework involved a discriminator and a generator working in opposition to refine the quality and authenticity of generated images. The discriminator's objective was to accurately classify real and generated images, while the generator aimed to produce images that the discriminator would mistake as real. This adversarial process was conditioned on auxiliary information, enabling the generation of images tailored to specific contexts or attributes.

The mathematical expressions for the discriminator and generator objectives provided a rigorous foundation for implementing and optimizing the adversarial model. Through iterative training, the model demonstrated an increasing proficiency in generating images that closely mimic real data in terms of both appearance and diversity. This was evidenced by the progression of generated images over the course of training, starting from basic representations to more complex and detailed visual outputs.

Advancements in generative models should focus on improving diversity in generated images and integrating GANs with other architectures for enhanced realism. These principles can be applied beyond image generation, extending to domains like video generation and drug discovery. Ethical considerations must be addressed to ensure responsible use. Continued refinement and exploration of applications hold promise for significant contributions to technology and society.

REFERENCES

[1] T. Wang, T. Zhang and B. Lovell, "Faces à la Carte: Text-to-Face Generation via Attribute Disentanglement," 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2021, pp. 3379-3387, doi: 10.1109/WACV48630.2021.00342.

[2] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding," submitted to arXiv on May 23, 2022. [Online]. Available: https://arxiv.org/abs/2205.11487

[3] X. Wang, T. Qiao, J. Zhu, A. Hanjalic and O. Scharenborg, "Generating Images From Spoken Descriptions," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 850-865, 2021, doi: 10.1109/TASLP.2021.3053391.

[4] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion," submitted to arXiv on Aug. 2, 2022. [Online]. Available: https://arxiv.org/abs/2208.01618

[5] M. B. Bejiga, F. Melgani and A. Vascotto, "Retro-Remote Sensing: Generating Images From Ancient Texts," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 12, no. 3, pp. 950-960, March2019, doi:10.1109/JSTARS.2019.2895693.

[6] S. Li, Y. Zhang, M. Huang, H. Wu and W. Cai, "Building and Using a Supply Chain Knowledge Graph applied to the rail transit industry," 2021 5th Asian Conference on Artificial Intelligence Technology (ACAIT), Haikou, China, 2021, pp. 742-746,doi:10.1109/ACAIT53529.2021.9731237.

[7] H. Zhang et al., "StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 8, pp. 1947-1962, 1 Aug. 2019, doi: 10.1109/TPAMI.2018.2856256.

[8] X. Hou, X. Zhang, Y. Li and L. Shen, "TextFace: Text-to-Style Mapping Based Face Generation and Manipulation," in IEEE Transactions on Multimedia, vol. 25, pp.3409-3419,2023, doi:10.1109/TMM.2022.3160360.

[9] Liu, Sean. (2019). Text-To-Image Generation. 10.13140/RG.2.2.14173.56801

[10] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, "StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks," submitted to arXiv on Dec. 10, 2016, last revised Aug. 5, 2017 (version 2). [Online]. Available: https://arxiv.org/abs/1612.03242

[11] X. Chen, L. Qing, X. He, X. Luo, and Y. Xu, "FTGAN: A Fully-trained Generative Adversarial Networks for Text to Face Generation," submitted to arXiv on Apr. 11, 2019. [Online]. Available: https://arxiv.org/abs/1904.05729

[12] 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," submitted to arXiv on Nov. 26, 2019. [Online]. Available: https://arxiv.org/abs/1911.11378

[13] D. Valevski, D. Wasserman, Y. Matias, and Y. Leviathan, "Face0: Instantaneously Conditioning a Text-to-Image Model on a Face," submitted to arXiv on Jun. 11, 2023. [Online].Available: https://arxiv.org/abs/2306.06638

[14] Optimal text-to-image synthesis model for generating portrait images using generative adversarial network techniques- http://doi.org/10.11591/ijeecs.v25.i2.pp972-979

[15] M. Z. Khan et al., "A Realistic Image Generation of Face From Text Description Using the Fully Trained Generative Adversarial Networks," in IEEE Access, vol. 9, pp. 1250-1260, 2021, doi: 10.1109/ACCESS.2020.3015656.

[16] R. Yanagi, R. Togo, T. Ogawa and M. Haseyama, "Query is GAN: Scene Retrieval With Attentional Text-to-Image Generative Adversarial Network," in IEEE Access, vol. 7, pp. 153183-153193, 2019,doi:10.1109/ACCESS.2019.2947409.

[17] R. Bayoumi, M. Alfonse and A. -B. M. Salem, "An Intelligent Hybrid Text-To-Image Synthesis Model for Generating Realistic Human Faces," 2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS), Cairo, Egypt, 2021, pp. 172-176,doi:10.1109/ICICIS52592.2021.9694194.

[18] H. Tan, X. Liu, M. Liu, B. Yin and X. Li, "KT-GAN: Knowledge-Transfer Generative Adversarial Network for Text-to-Image Synthesis," in IEEE Transactions on Image Processing, vol. 30, pp. 1275-1290, 2021, doi: 10.1109/TIP.2020.3026728.

[19] Xue, Yuting & Zhou, Heng & Ding, Yuxuan & Shan, Xiao. (2022). Adaptive Forgetting, Drafting and Comprehensive Guiding: Text-to-Image Synthesis with Hierarchical Generative Adversarial Networks. 273-285. 10.5121/csit.2022.120623.

[20] D. M. A. Ayanthi and S. Munasinghe, "Text-to-Face Generation with StyleGAN2," submitted to arXiv on May 25, 2022. [Online]. Available: https://arxiv.org/abs/2205.12512

[21] G. Yin, B. Liu, L. Sheng, N. Yu, X. Wang and J. Shao, "Semantics Disentangling for Text-To-Image Generation," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 2322-2331, doi: 10.1109/CVPR.2019.00243.

[22] R. Wadhawan, T. Drall, S. Singh and S. Chakraverty, "Multi-Attributed and Structured Text-to-Face Synthesis," 2020 IEEE International Conference on Technology, Engineering, Management for Societal impact using Marketing, Entrepreneurship and Talent (TEMSMET), Bengaluru, India, 2020, pp. 1-7, doi: 10.1109/TEMSMET51618.2020.9557583.

[23] A. Kushwaha, C. P and K. P. Singh, "Text to Face generation using Wasserstein stackGAN," 2022 IEEE 9th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), Prayagraj, India, 2022, pp. 1-7, doi: 10.1109/UPCON56432.2022.9986391.

[24] S. Jiang, Y. Shi and K. Cheng, "Text-to-Face Generation via Multi-Modal Attention Memory Network with Fine-Grained Feedback," 2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI), Chengdu, China, 2022, pp. 940-947, doi: 10.1109/PRAI55851.2022.9904267.

[25] R. Zhao and Z. Shi, "Text-to-Remote-Sensing-Image Generation with Structured Generative Adversarial Networks," in IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1-5, 2022, Art no. 8010005, doi: 10.1109/LGRS.2021.3068391.

[26] C. Zhang and Y. Peng, "Stacking VAE and GAN for Context-aware Text-to-Image Generation," 2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM), Xi'an, China, 2018, pp. 1-5, doi: 10.1109/BigMM.2018.8499439.