Text-to-Image Synthesis using DL Model

Submitted By

Jwal Shah

20BCE114

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

December 2023

Text-to-Image Synthesis using DL Model

Minor Project

Submitted in partial fulfillment of the requirements

for the degree of

Bachelor of Technology in Computer Science and Engineering

Submitted By

Jwal Shah

(20BCE114)

Guided By

Dr. Tejal Upadhyay

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

December 2023

Certificate

This is to certify that the minor project entitled **"Text-to-Image Synthesis using DL Model"** submitted by **Jwal Shah (20BCE114)**, towards the partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering, Nirma University, Ahmedabad, is the record of work carried out by him under my supervision and guidance. In my opinion, the submitted work has reached the level required for being accepted for examination. The results embodied in this minor project, to the best of my knowledge, haven't been submitted to any other university or institution for the award of any degree or diploma.

Dr. Tejal Upadhyay
Assistant Professor
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Dr Madhuri Bhavsar
Professor and Head,
CSE Department
Institute of Technology,
Nirma University, Ahmedabad

Statement of Originality

I, **Jwal Shah**, Roll. No. **20BCE114**, give an undertaking that the Minor Project entitled **"Text-to-Image Synthesis using DL Model"** submitted by me, towards the partial fulfillment of the requirements for the degree of Bachelor of Technology in **Computer Science and Engineering**, Nirma University, Ahmedabad, contains no material that has been awarded for any degree or diploma in any university or school in any territory to the best of my knowledge. It is the original work carried out by me and I give assurance that no attempt of plagiarism has been made. It contains no material that is previously published or written, except where reference has been made. I understand that in the event of any similarity found subsequently with any published work or any dissertation work elsewhere; it will result in severe disciplinary action.

Signature of Student
Date: 8|12|23
Place: Ahmedabad

Endorsed by
Tejal Upadhyay
(Signature of Guide)

Acknowledgements

I express my sincere appreciation to Mrs. Tejal Upadhyay for her invaluable guidance and unwavering support throughout the development of our minor project, titled "Text-to-Image Synthesis using DL Model" Mrs. Upadhyay expertise and mentorship played a crucial role in shaping the trajectory of our project.

Mrs. Upadhyay's contribution surpassed conventional guidance; she diligently assisted in formulating a comprehensive roadmap for the implementation of our project, offering a clear and structured path to success. Her adept teaching skills were instrumental in understanding complex concepts related to path planning algorithms, enabling me to navigate the intricacies of our chosen subject matter.

I am particularly thankful for the wealth of innovative ideas she shared to overcome challenges encountered during the project. Mrs. Upadhyay's insightful perspectives and problem-solving approach significantly enriched the project's outcomes. Learning under Dr. Tejal Upadhyay's guidance has been a transformative experience, enhancing various facets of my academic journey. I genuinely appreciate her dedication and mentorship, which have played a pivotal role in my growth and success.

In conclusion, I extend my gratitude to Dr. Tejal Upadhyay for being an inspiring guide and making a profound impact on my academic endeavors.

- Jwal Shah

20BCE114

Abstract

Text-to-image synthesis refers to converting textual features into pixels, which requires a full understanding of the connection between the natural language text and visual features. Text-to-image creation, or the generation of lifelike images from textual descriptions, is a tough task. The purpose of this endeavor is to help machines understand, interpret, and translate natural language descriptions into visual representations. Approaches based on deep learning have made substantial progress in creating realistic images from text-based descriptions. This was accomplished by employing generative adversarial networks (GANs) and attention techniques. These methods have yielded encouraging results in terms of producing images of excellent quality that are equally semantically and visually meaningful. However, other hurdles remain, such as increasing the diversity and resolution of generated images and dealing with increasingly sophisticated textual descriptions. This work focuses on the implementation of stackGAN for text-to-image generation, which creates photo-realistic images with a resolution of 256 x 256. In two steps, we decompose the problem of text-to-image production. The first stage of the model outlines the primitive features of the object, such as shape and color, and provides a low-quality image. The stage two generator generates photo-realistic images based on the text description provided as input from the results of the stage one generator. Conditional augmentation is also employed to boost the generated images' diversity, stabilizing the training of conditional GAN.

Abbreviations

GANs	Generative Adversarial Networks.
AI	Artificial Intelligence.
ML	Machine Learning.
DL	Deep Learning
T2I	Text to image.
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
ML	Machine Learning
AI	Artificial Intelligence
GPU	Graphics Processing Unit
ReLU	Rectified Linear Unit
Loss Func.	Loss Function
FC	Fully Connected (layer)
CNN-RNN	Combination of CNN and RNN

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The article addresses the difficulty of translating text to high-resolution images and emphasises the shortcomings of standard Generative Adversarial Networks (GANs) for this purpose. It presents StackGAN, a novel paradigm that uses a two-stage generation process to address these issues.

StackGAN uses a conditional GAN in the first stage to create a low-resolution image from a text description. Then, to create a high-resolution image, the second stage makes use of both the verbal description and the low-resolution image created in the first stage. This two-step process produces realistic and varied graphics by enabling the model to capture the global structure as well as the fine-grained details of the images.

A thorough analysis of StackGAN's technical aspects, including its architecture, training methods, and assessment procedures, is given in this study. It goes over the benefits and drawbacks of the model and contrasts it with other cutting-edge methods for generating text from images. Furthermore, StackGAN's potential uses in the creative industries—such as design, art, and entertainment—are investigated. The paper ends with several possible directions for further study in the field of text-to-image synthesis. Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. Databases, data warehouses, the Internet, other information repositories, and data that is dynamically fed into the system are examples of data sources.

Chapter 2

Related Work

2.1 Techniques used in previous work :-

2.1.1 Generative adversarial text to image synthesis

A unique deep architecture and GAN formulation is presented in [1] with the goal of bridging the gap between text and image modelling advancements. The translation of human-written descriptions into realistic images of birds and flowers is the main emphasis of the work. The suggested approach successfully combines tenable visual interpretations, proving that style and content can be separated, and it also successfully transfers the background and posture of the bird from query images to text descriptions. Text-to-image synthesis on CUB is enhanced by the addition of a manifold interpolation regularise, demonstrating generalizability to produce images with many objects and varying backgrounds on the MS-COCO dataset. More text formats and higher resolution image scaling are the goals of future work on the model.

2.1.2 A survey into ai text-to-image generation in the era of large model

The development of Text-to-Image (TTI) generation models is examined in [2], which traces their roots from autoregressive Transformers and diffusion models to Generative Adversarial Networks (GANs). It emphasises how important diffusion models are to image synthesis, emphasising the need to integrate large language models and scale up model size for better results. The paper highlights the groundbreaking influence of TTI models in AI provided Content (AIGC), demonstrating their capacity to produce content

that is indistinguishable from content provided by humans. The study ends with a picture of how TTI models might be used in the future, including the production of videos and 3D content, which suggests important developments in the content generating field.

2.1.3 Stackgan: Text to photo realistic image synthesis with stacked generative adversarial networks.

The research article [3] addresses the problem of producing high-resolution (256x256) photo-realistic images from text descriptions by introducing Stacked Generative Adversarial Networks (StackGAN) with Conditioning Augmentation. The suggested approach uses a two-step procedure: Stage-I GAN uses text to generate crude shapes and colours, while Stage-II GAN improves and refines details to produce realistic-looking graphics. Conditional-GAN training is stabilised and diversity is improved using the Conditioning Augmentation approach. Extensive experiments reveal notable gains over current models, resulting in better quality images with more details and realism, indicating the promise of the technology for computer-aided design and photo-editing applications.

2.1.4 Stackgan++: Realistic image synthesis with stacked generative adversarial networks.

Stacked Generative Adversarial Networks (StackGANs) are presented in [4] as a potential solution to the problems associated with producing high-quality, high-resolution photographs. The StackGAN-v1 and StackGAN-v2 designs that have been suggested deal with generative tasks and text-to-image synthesis. While StackGAN-v2 uses a multi-stage structure with several generators and discriminators, StackGAN-v1 uses a two-stage technique. The outcomes show that these stacked GANs perform noticeably better than the most advanced techniques, offering more consistent training and producing images that are almost lifelike in both conditional and unconditional settings.

2.1.5 Text-to-image generation using multi-instance stackgan

A Stacked Generative Adversarial Networks (StackGAN) model that is improved is presented in [5] to tackle the difficult problem of producing high-resolution images from text descriptions that have many instances in various categories. Based on the semantics of the input language, the model shows that it can compose large scenes with many items. The study underscores the significance of deep networks in text-to-image tasks,

demonstrates the efficacy of the StackGAN model, and proposes directions for future development, with a focus on longer training for Stage II GAN, investigating alternative architectural configurations, fine-tuning hyperparameters, and optimising preprocessing to boost model performance.

Chapter 3

Stacked Generative Adversarial Network

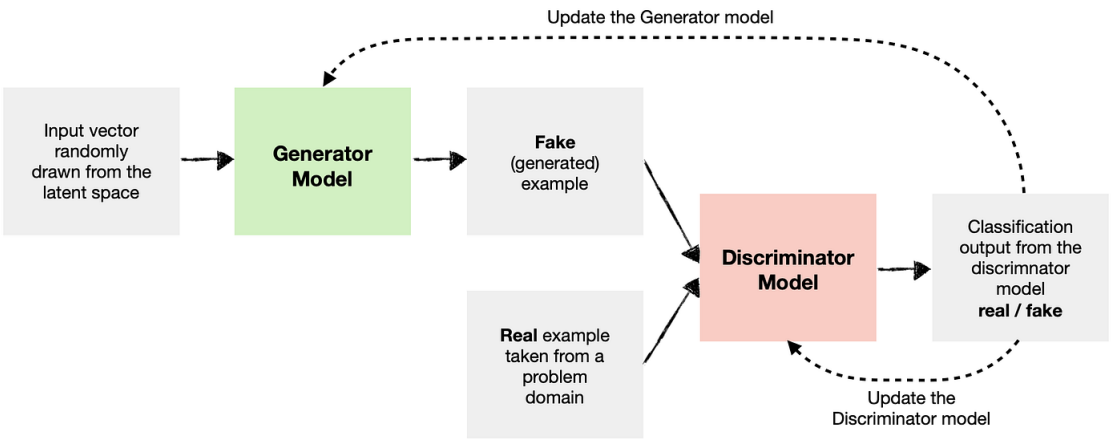

Figure 3.1: Generator and Discriminator Model

3.1 About

3.1.1 Groundwork

A discriminator and a generator playing a min-max game make up a Generative Adversarial Network (GAN). The discriminator (D) seeks to discern between actual and produced data, while the generator (G) produces superior synthetic data to trick the discriminator. An extra condition (c) is added to conditional GANs. The antagonistic relationship is outlined by the min-max value function, usually represented as V(D, G). While G wants to maximise the realism of generated data in order to trick D, D wants to minimise the classification error between genuine and generated data. During the whole

GAN training process, these goals are balanced until a Nash equilibrium is attained, at which point generated data can no longer be distinguished from real data.

3.1.2 Conditioning Augmentation

Using pre-trained word2vec models with transfer learning, text descriptions are embedded. By adding more training data to the generator, conditioning augmentation helps with the shortage of text data. By using this method, robustness to little changes in the conditioning manifold is improved. The generator's target is extended with a regularisation period to guarantee smoothness throughout training and avoid overfitting. A Gaussian distribution is used in the newly developed KL divergence function to represent the variability in text-to-image translation. The unpredictability in Conditioning Augmentation is influenced by the mean and covariance matrix parameters. In text-to-image translation, this unpredictability represents the various locations and appearances connected to a particular sentence. An example of the KL divergence function is shown.

3.2 Stages

3.2.1 Stage-1

Stage I involves using conditioned text descriptions to create a low-resolution image that highlights the colour and shape of the object. There are two phases to the endeavour, the first of which focuses on producing low-resolution photographs. For text descriptions, pre-trained encoders provide the embeddings t. To capture fluctuations in the meaning of t, samples of the Gaussian conditioning variables c0 are taken. Optimising LG0 and minimising LD0, the Stage I GAN trains generator G0 and discriminator D0 in turn. These goals are balanced by the regularisation term . Ilow-res stands for low-resolution photos taken in Stage I, when intrinsic constraints may cause distortions. Acquiring a basic knowledge of the object's visual characteristics is the main objective of Stage I.

3.2.2 Stage-2

The forms and details of objects are not clearly seen in the first low-resolution photos from Stage-I GAN. By using text embeddings and low-resolution images to rectify faults, Stage-II GAN improves on the outcomes of Stage-I. The process of creating high-resolution images is represented by the equation Ihigh-res = Stage2(Ilow-res, c). Shape,

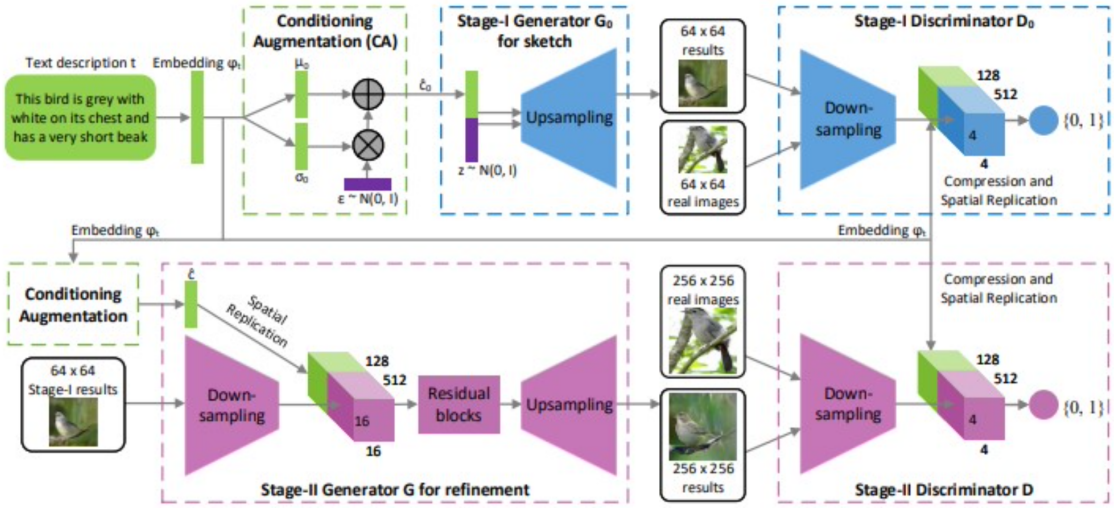

Figure 3.2: The architecture of the proposed stack GAN with stage 1 and stage 2 working flow.

Table 3.1: Discriminator Output

Input to Discriminator	**Output**
Real Image, Real Caption	1
Wrong Image, Real Caption	0
False Image, Real caption	0

colour, and finer object characteristics are captured in the final photos through an iterative refining process that guarantees close alignment with the input word descriptions. The residual block encoder-decoder network used by the Stage-2 generator creates a conditioning vector by means of text embedding. In order to enhance alignment between photos and conditioning text, the Stage-2 discriminator uses a matching-aware technique and can handle higher image sizes. In order to accomplish StackGAN's text-to-image generation, Stage I and II work together to provide a preliminary understanding and continuously improve the output to provide high-quality, photo-realistic results.

3.3 Different models in comparison with StackGAN

A cutting-edge GAN framework known as StyleGAN [6] is capable of producing images as high as 1024x1024 in resolution, among other exceptional quality images [14]. Creating a stack of layers with the intention of starting with a low-resolution image (starting at

2x2) and progressively increasing the resolution to higher levels is the goal. StyleGAN is used to generate high-resolution images with a range of properties by feeding it a random vector or noise. StackGAN [7] receives a low-resolution image and a written description as input, and uses them to produce high-resolution images that are conditioned on the text. Compared to StyleGAN, which can produce images with a resolution of up to 1024x1024, StackGAN produces images with a lower resolution. StyleGAN's progressive growth technique teaches the generator and discriminator by progressively increasing the resolution of the generated images.

It also uses a mapping network to translate the input noise vector to an intermediate latent space, which produces the final image. StackGAN uses a two-stage generator architecture to produce graphics in two steps. The first stage generates a low-quality image based on the input text, which is subsequently enhanced to a higher resolution in the second stage. StackGAN creates multi-modal graphics with a range of features and styles by conditioning on the input text. Similar to this, StyleGAN is a multimodal image generation model capable of generating images with a wide range of features and high resolution. Both require expensive computing power and a significant investment in time and computer resources for training. However, it is more stable during training than other GAN models, and it regularly generates high-quality, featured images. StackGAN may, however, suffer from overfitting or mode collapse during training. Unlike StyleGAN, which can generate high-quality images across multiple domains, StackGAN is a more specialised model designed specifically to generate text-conditioned images.

Chapter 4

Experiments

We carry out extensive quantitative and qualitative analyses to validate our methodology. This work compares GAWWN [6] and GAN-INT-CLS [5], two state-of-the-art methods for generating visuals from text descriptions. The code provided by the authors of the two methodologies that are being compared is used to create the findings. In addition, we create a number of baseline models to examine the general architecture and key elements of our suggested StackGAN. To evaluate the effectiveness of the proposed layered architecture with Conditioning Augmentation, we trained the Stage-1 GAN to generate 64x64 and 256x256 images for the first baseline right away. Next, we tweaked our StackGAN design to investigate whether our approach could yield higher-resolution images with greater quality. The model was trained to produce 256x256 and 128x128 images in order to accomplish this. Additionally, we look into how text input is used in the first two phases of StackGAN.

4.1 Dataset Used

200 bird species are represented by 11,788 photos in CUB [13]. We crop every picture as part of the preparation procedure to make sure that the bounding boxes of birds have object-image size ratios greater than 75/100, since 80 percent of the birds in this dataset had object-image size ratios of less than 0.5 [13]. 8,189 photos of flowers from Oxford-102 are organised into 102 groups. We also assess the generalizability of our method on the more difficult dataset, MS COCO. The MS COCO dataset differs from the CUB and Oxford-102 databases in that it contains images with a range of backgrounds and objects. In the COCO dataset, each image has five descriptions; in the CUB and Oxford-

102 datasets, each photo has ten descriptions. There are two sets of photos in it: 40,000 images for validation and 80,000 images for training. We divide the CUB and Oxford-102 training and test sets by class, and we employ the COCO training and validation sets according to the experimental setup.

4.2 Evaluation Metrices

Given the challenges associated with evaluating the performance of generative models, we have selected the inception score as our recommended numerical assessment method for quantitative evaluation. A single generated sample is represented by x, and the expected label of the inception model is represented by y in [8]. This statistic is based on the idea that effective models should produce a variety of interesting and useful visualisations. In KL, there should be a considerable difference between the marginal distribution p(y) and the conditional distribution p(y—x). We employ the pre-trained COCO dataset Inception model right away in our research. We optimise an Inception model for both the CUB and Oxford-102 fine-grained datasets. We assess this metric on a large sample size (30k randomly selected data) for every model, as stated in the paper. Although the inception score has a strong correlation with individuals' assessments of the samples' visual quality, it is unable to determine if the generated images are suitably conditioned by the given text descriptions. As a result, we evaluate people too. For the CUB and Oxford-102 exam sets, 50 text descriptions will be randomly selected for each class [9]. From the COCO dataset's validation set, 4,000 text descriptions are chosen at random. From each model, five images are produced for every syllable. The same written descriptions are provided to ten users (none of whom are writers), and they are asked to assess the results using various techniques. In order to assess all contrasting approaches, we calculate the average ranks by human users.

Figure 4.1: Our StackGAN's example results are compared to those of GAWWN and GAN-INT-CLS

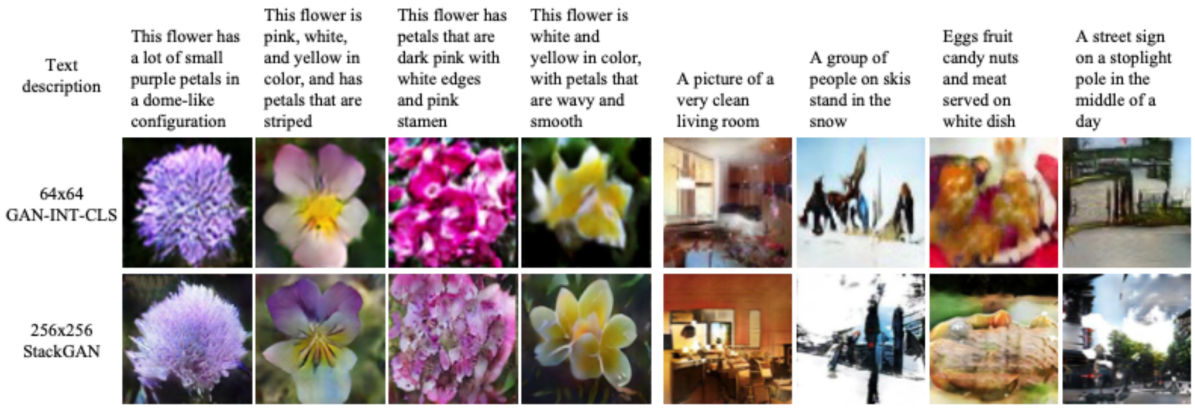

Figure 4.2: Our StackGAN's performance was evaluated and compared to that of GAN-INT-CLS, which is also conditioned on textual descriptions from the Oxford-102 text set and COCO validation set.

Chapter 5

Results

StackGAN is compared with current text-to-image techniques in the paper using the CUB, Oxford-102, and COCO datasets. StackGAN outperforms other models in terms of both average human score rank and inception. Compared to GAN-INT-CLS, it significantly raises the inception score on the CUB and Oxford-102 datasets by 28.47% and 20.30%, respectively. StackGAN outperforms GAWWN, which needs further restrictions for higher scores, in producing 256x256 photorealistic images only from text descriptions. The outcomes show that while Stage-2 GAN fine-tunes details to produce more realistic images, Stage-1 GAN captures broad forms and colours but lacks clarity. StackGAN's realism is enhanced by its capacity to improve backdrops in Stage-2 and rectify faults in Stage-1. Rather than being a result of memorization, the performance is linked to grasping intricate language-image links. StackGAN extracts visual features using the Stage-2 discriminator and displays variations from the training set while preserving common characteristics. All things considered, StackGAN does a great job producing varied and high-quality images from textual descriptions, demonstrating its efficacy in complex language-image understanding.

Table 5.1: Results

Metric	**Dataset**	**GAN-INT-CLS**	**GAWWN**	**Our StackGAN**
Inception Score	CUB	2.88±.04	3.62±.07	3.70±.04
	Oxford	2.66±.03	/	3.20±.01
	COCO	7.88±.07	/	8.45±.03
Human rank	CUB	2.81±.03	1.99±.04	1.37±.02
	Oxford	1.87±.03	/	1.13±.03
	COCO	1.89±.04	/	1.11±.03

Bibliography

[1] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *International conference on machine learning*, pp. 1060–1069, PMLR, 2016.

[2] F. Bie, Y. Yang, Z. Zhou, A. Ghanem, M. Zhang, Z. Yao, X. Wu, C. Holmes, P. Golnari, D. A. Clifton, *et al.*, "Renaissance: A survey into ai text-to-image generation in the era of large model," *arXiv preprint arXiv:2309.00810*, 2023.

[3] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 5907–5915, 2017.

[4] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan++: Realistic image synthesis with stacked generative adversarial networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1947–1962, 2018.

[5] A. Fu and Y. Hou, "Text-to-image generation using multi-instance stackgan," *Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Sprint 2017*, pp. 225–231, 2017.

[6] Y. Imrana, Y. Xiang, L. Ali, and Z. Abdul-Rauf, "A bidirectional lstm deep learning approach for intrusion detection," *Expert Systems with Applications*, vol. 185, p. 115524, 07 2021.

[7] M. Rohith, L. Pallavi, K. Shirisha, M. Sanjay, and V. S. Priya, "Image generation based on text using bert and gan model," in *2023 International Conference on Compu-*

tational Intelligence, Communication Technology and Networking (CICTN), pp. 214–218, 2023.

[8] Z. Yi, Z. Chen, H. Cai, W. Mao, M. Gong, and H. Zhang, "Bsd-gan: Branched generative adversarial network for scale-disentangled representation learning and image synthesis," *IEEE Transactions on Image Processing*, vol. 29, pp. 9073–9083, 2020.

[9] L. Xiaolin and G. Yuwei, "Research on text to image based on generative adversarial network," in *2020 2nd International Conference on Information Technology and Computer Application (ITCA)*, pp. 330–334, 2020.