TEXT-TO-IMAGE GENERATION USING GAN.

MSc Research Project Artificial Intelligence

Ganesh Kesham

Student ID: x22170812

School of Computing

National College of Ireland

Supervisor: Muslim Jameel Syed

National College of Ireland

MSc Project Submission Sheet School of Computing

Student Name:	Ganesh Kesham		
Student ID:	X22170812		
Programme:	Msc in Artificial Intelligence	**Year:**	2023
Module:	Msc Research Practicum		
Lecturer:	Muslim Jameel Syed		
Submission Due Date:	31/01/2024		
Project Title:	TEXT-TO-IMAGE GENERATION USING GAN.		
Word Count:	**Word Count: 5871**	**Page Count : 29**	

Abstract

This research delves into the pursuit of generating realistic images from textual descriptions, a compelling yet challenging task within current AI systems. While existing technologies fall short of this goal, recent advancements in recurrent neural networks have demonstrated proficiency in learning discriminative text features. Additionally, deep convolutional generative adversarial networks (GANs) have shown promise in generating highly realistic images across specific categories like faces, album covers, and interiors. In this study, I introduce a novel deep architecture and GAN formulation aimed at bridging these text and image modelling advancements. Here approach is centered on translating textual concepts into vivid visual representations, effectively converting characters into pixelated images. Through rigorous experimentation, we showcase the capability of our model to produce credible images of birds and flowers from intricate textual

descriptions. This work represents a significant step toward achieving the synthesis of detailed images solely from text, offering insights into the convergence of text and image modelling within the realm of artificial intelligence.

Keywords: Text-to-Image Generation, Generative Adversarial Networks (GANs), Image Synthesis

Chapter 1 Introduction

1.1 Background

The confluence of artificial intelligence (AI) in natural language processing (NLP) and computer vision in recent years has resulted in significant advances in the field of text-to-image creation. This emerging discipline aims to bridge the gap between verbal descriptions and visual representations, culminating in the creation of realistic visuals only from comprehensive word inputs. The capacity to transform textual semantics into visually appealing visuals has enormous potential in a variety of disciplines, including design, content development, and augmented reality applications (Ramesh et al., 2021). The automatic production of realistic images from written descriptions has a wide range of possible applications, including image editing, video games, and accessibility. This research has a goal of text-to-image creation is to produce realistic visuals that match the provided text in terms of semantic consistency (Huang et al., 2022). The requirement for many high-quality text-image pairs is one of the main obstacles in training text-to-image generating algorithms (Zhou et al., 2022). Thus, in order to address this problem, GAN networks will be used, as stated by (Bertrand and Azizi, 2022), which is exactly what we will do in our research.

1.1.1 Historical Evolution and Milestones

Traditional AI systems have excelled in specific fields such as natural language understanding and image identification (Khan et al., 2023). However, synergy across these fields remains a difficult frontier. The hunt for coherent and contextually meaningful images from textual descriptions has piqued the interest of many researchers. This approach combines subtle textual comprehension with elaborate visual interpretation—a union that necessitates novel neural network designs and advanced learning processes.

1.2 Aim of the study

The fundamental goal of this research is to go ahead in the field of artificial intelligence by concentrating on the creation of a revolutionary method for producing realistic visuals exclusively from precise textual descriptions. In order to achieve this goal, the research will take advantage of recent advances in recurrent neural networks (RNNs) for learning discriminative text features, as well as the capabilities of deep convolutional generative adversarial networks (GANs) for producing highly realistic images within specific categories. The project aims to bridge the gap between text and image modelling by incorporating these advances, culminating in the development of a robust deep architecture and GAN formulation. The ultimate aim is to demonstrate the model's ability to translate complex textual notions like birds and flowers into visually appealing and credible image outputs, therefore greatly advancing the area of text-to-image creation in artificial intelligence.

1.3 Research Objectives

The research objectives of this report are:

1. The primary goal is to create a novel deep architecture that combines recurrent neural networks (RNNs) with convolutional generative adversarial networks (GANs) to bridge the gap between textual descriptions and visually appealing image outputs.
2. The second goal is to train and optimise GAN networks, which entails holding lengthy training sessions that last for around 24 hours and 450 epochs to optimise the GAN framework's Generator and Discriminator networks.
3. The third goal of this research is to analyse the achieved losses for both the Generator and Discriminator networks to thoroughly assess the text-to-image generating model's performance.

1.4 Research Questions

The research questions for this report are:

1. The first research challenge concerns how to use sophisticated neural network designs, such as convolutional generative adversarial networks (GANs) and recurrent neural networks (RNNs), to successfully construct a system that transforms textual descriptions into realistic visuals.
2. The second research question is what training strategies and optimization techniques are most effective in achieving convergence and enhancing the performance of the text-to-image generation model over extended epochs and training durations?
3. In what ways can the outputs of the text-to-image generation model be comprehensively evaluated, both quantitatively and qualitatively, to assess the fidelity, diversity, and realism of the generated images?

1.5 Research Gaps

The research gaps in this report lie in:

1. Semantic Understanding in Image Generation
2. Fine-Grained Image Realism

Chapter 2 Literature Review

2.1 Text to Image Generation using RNN

This section of the literature review chapter converges on the basic goals of achieving realistic image synthesis as well as semantic alignment between text and images through innovative models and attention mechanisms, indicating a shared pursuit of enhancing the intersection of computer vision and natural language understanding. (Sharma et al., 2018) underline the relevance of including conversation in addition to captions to give deeper context for image synthesis tasks. (Dong et al., 2017) proposed combining text-to-image and image-to-text synthesis to boost performance while demonstrating transfer learning capabilities. MirrorGAN, developed by (Qiao et al., 2019), focuses on both visual realism and semantic consistency by using attention processes and regeneration modules. (Zia et al., 2020) created a stable and tractable caption-based image-generating model that prioritized word-to-pixel relationships. (He et al., 2017) emphasized the importance of creating natural language descriptions from images and the implications for a variety of applications.

Text-to-image production approaches have seen a spike in interest and investigation in recent years, reflecting their importance across multiple academic disciplines and practical applications such as image, art development, and assisting visually impaired persons. Both (Chen et al., 2018) and (Ramzan et al., 2022) address the issue of semantic consistency in image production from text by presenting deep learning architectures—RC-GAN and a new deep model, respectively—to create more realistic image synthesis. (Zhuge et al., 2018) and (Xu et al., 2018) are interested in establishing attentional mechanisms inside

Generative Adversarial Networks (GANs) to enable fine-grained text-to-image synthesis by attending to regions depending on text cues. (Reed et al., 2016) and (Xu et al., 2018) highlight the merging of textual and image modalities, offering unique structures aimed at translating visual notions from written descriptions to pixel representations. These works attempt to improve image synthesis from text by utilizing deep learning, attention processes, and GANs, with the goal of achieving more realistic and semantically matched image production from textual descriptions. The attempt to bridge the gap between verbal descriptions and image production, striving for semantic coherence and visual realism, is a recurrent thread throughout these works.

Table 2.1: Table for Text to Image Generation using RNN

Study	Focus	Proposed Approach	Challenges Addressed	Results/Evaluation
Sharma et al. (2018)	Enhancing image synthesis using dialogue alongside captions	Integration of dialogue for richer context in image synthesis	Insufficient caption information, object-word correspondence	Improved Inception Score, and quality in MS COCO dataset
Dong et al. (2017)	Integrating text-to-image and image-to-text synthesis	Fusion of synthesis methods for performance improvement	Multi-category image complexity, transfer learning	Superior performance in multi-category image gen, transfer learning capabilities
Qiao et al. (2019)	Visual realism and semantic consistency	MirrorGAN framework, attention mechanisms, regeneration modules	Realism, semantic coherence, attention mechanisms	Outperformed other methods in realism, semantic consistency
Zia et al. (2020)	Stable, tractable caption-based image generation	Attention-based encoder, autoregressive decoder	Intractable inference, instability in training	Better quality images, leveraging word-to-pixel dependencies
He et al. (2017)	Descriptive natural language from images	Overview of visual captioning advancements and applications	Interdisciplinary challenges, future breakthroughs	Impacts on research, industry deployment, future directions

Chen et al. (2018)	Semantic consistency in image generation	RC-GAN architecture for more realistic image synthesis	Semantic consistency in image synthesis	Improved realism in image synthesis
Zhuge et al. (2018)	Attentional mechanisms in GANs	Attentional GAN model for fine-grained text-to-image synthesis	Fine-grained synthesis, attention mechanisms	Improved fine-grained text-to-image synthesis
Ramzan et al. (2022)	Realistic image synthesis from text	Novel deep learning model for semantic consistent image generation	Semantic consistency, realistic synthesis	Enhanced realism in image synthesis, semantic alignment
Reed et al. (2016)	Fusion of textual and image modalities	Introducing novel architectures for text-to-image pixel translation	Merging textual and visual concepts, model architecture	Improved translation from textual descriptions to images
Xu et al. (2018)	Attention-driven text-to-image synthesis	AttnGAN architecture with attention mechanisms for image generation	Attention-driven synthesis, semantic consistency	Enhanced attention-based synthesis, semantic alignment

2.1 Text to Image Generation using GAN

Image synthesis from text descriptions is a challenging yet developing area of computer vision research. To handle this challenge, several strategies have been presented, each with its distinct contributions and approaches. (Liao et al., 2022) proposed the Semantic-Spatial Aware GAN framework, which focuses on matching produced images with text descriptions at both the holistic and fine-grained levels, solving regional consistency issues. Similarly, (Zhu et al., 2019) presented the Dynamic Memory Generative Adversarial Network (DM-GAN) to improve ambiguous initial image contents, focusing on the dynamic selection of critical text information for accurate image production. In their FuseDream pipeline, (Liu et al., 2021) coupled CLIP representations with off-the-shelf GANs in an original way, emphasizing optimization in the GAN space using creative ways to create varied, high-quality images. (Ruan et al., 2021) addressed the aspect information difficulty by developing the Dynamic Aspect-aware GAN (DAE-GAN), which incorporates complete text representations at several granularities and refines images at the aspect and local levels. Moreover, (Wang et al., 2021) introduced the Cycle-consistent Inverse GAN (CI-GAN), which unifies text-guided image creation and manipulation tasks through cycle-consistency training and latent space semantics exploration.

(Zhang et al., 2018) suggested a context-aware technique that separates foreground and background using a conditional VAE-GAN structure, hence improving text-image alignment. (Tan et al., 2022) used semantic disentanglement and distribution normalization modules to improve image creation from text and introduced

distribution regularization in GANs. (Sawant et al., 2021) concentrated on criminal face creation, using GANs to turn textual descriptions of facial characteristics into realistic human features. (Tao et al., 2022) introduced DF-GAN, which addresses entanglement difficulties in generator designs while also increasing semantic consistency and deepening text-image fusion for improved image synthesis. (Ding et al., 2021) presented CogView, a text-to-image synthesis system that combines a 4-billion-parameter Transformer with a VQ-VAE tokenizer, displaying adaptability across several downstream applications and delivering state-of-the-art performance in text-to-image synthesis.

Table 2.2: Table for Text to Image Generation using GAN

Approach	Main Focus	Key Techniques/Modules	Datasets	Contributions
Liao et al. (2022)	Holistic and fine-grained alignment	Semantic-Spatial Aware Blocks	COCO, CUB	Regional consistency, semantic alignment
Zhu et al. (2019)	Refinement of initial images	Dynamic Memory Module	Caltech-UCSD Birds, COCO	Dynamic selection of crucial text information
Liu et al. (2021)	Diverse high-quality image generation	CLIP+GAN FusionDream Pipeline	MS COCO	Optimization in GAN space, diverse image generation
Ruan et al. (2021)	Incorporating aspect-level information	Aspect-aware Dynamic Re-drawer (ADR)	CUB-200, COCO	Aspect-level text representation, local refinement
Wang et al. (2021)	Unified text-guided generation/manipulation	Cycle-consistent Inverse GAN (CI-GAN)	Recipe1M, CUB	Unified framework for image synthesis & manipulation
Zhang et al. (2018)	Context-aware text-to-image generation	Context-aware conditional VAE, conditional GAN	Widely-used datasets	Separating foreground-background, improved alignment
Tan et al. (2022)	Distribution regularization in GANs	Semantic disentangling, distribution normalization	Public datasets	Improved image generation, semantic consistency
Sawant et al. (2021)	Criminal face generation	GANs for facial traits conversion	Specific criminal datasets	Realistic human face synthesis from textual traits
Tao et al. (2022)	DF-GAN: Simplified yet efficient text-to-image	One-stage backbone, Target-Aware Discriminator, Deep fusion block	Standard datasets	Improved architecture, semantic consistency

Ding et al. (2021)	CogView: Transformer-based text-to-image	4-billion-parameter Transformer, VQ-VAE tokenizer	Blurred MS COCO dataset	Versatility in downstream tasks, state-of-the-art FID

Chapter 3 Research Methodology

3.1 CRISP-DM Methodology

The Cross-Industry Standard Process for Data Mining, or CRISP-DM, is a widely accepted technique that provides an organized and complete framework for carrying out data-centric initiatives. It is divided into five stages: business comprehension, data comprehension, data preparation, modelling and evaluation. CRISP-DM acts as a guiding structure throughout the project lifecycle in the context of this study on text-to-image creation utilizing GANs. It begins with understanding the technology's main objectives and prospective applications—bridging word descriptions to visual representations—which aligns with the Business Understanding phase. Following this, the Data Understanding step explores and comprehends the flower's dataset, verifying its eligibility for training the GAN model. The Data Preparation phase includes the preparation operations required to convert image and text data into a model-compatible format. The modelling phase is concerned with the design and growth of the GAN architecture, whereas the evaluation phase examines the model's effectiveness in producing realistic images from text descriptions. CRISP-DM was utilised in this report, which contains various stages, which are as follows:

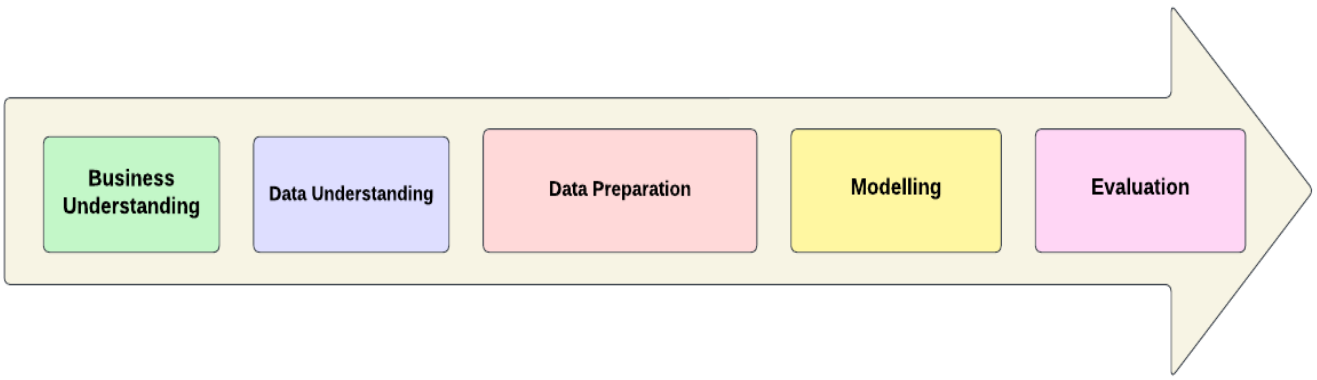

Figure 3.1: CRISP-DM Flow Diagram

1. **Business Understanding:** In the context of designing a text-to-image generation model utilising GANs, the business understanding phase entails knowing the technology's main aims and prospective applications. The major objective is to bridge the gap between verbal descriptions and visual representations by automating the generation of realistic graphics from comprehensive text. This breakthrough has ramifications in a variety of industries, including e-commerce, design, and content development, by allowing images to be generated based on verbal descriptions, improving user experience, enabling quick prototyping, and accelerating the creation of visual content. Understanding the possibilities and limits of this technology is critical for realising its promise in a variety of commercial areas, generating innovation, and producing value by computerising of image creation from textual input.
2. **Data Understanding:** Within the CRISP-DM architecture, the data understanding step entails a thorough examination of the flower dataset utilised in this text-to-image production project. This dataset contains 102 floral image categories, each with varied amounts ranging from 40 to 258 images per category, for a total of almost 7000 image-description pairings. Understanding the dataset structure, content distribution, and textual annotations is critical. Preprocessing steps include converting images into NumPy arrays based on predetermined pixel sizes, storing these arrays, and translating image descriptions into embeddings, subsequently saved in a CSV file. This phase focuses on gaining insights into the nature of the data, preparing it for subsequent model development, and ensuring its suitability for training the generative adversarial network (GAN) architecture aimed at generating plausible images from detailed text descriptions.
3. **Data Preparation:** Several critical actions are conducted during the data preparation phase of the CRISP-DM process for text-to-image creation to prepare the dataset for training the Generative Adversarial Network (GAN). This step entails preprocessing image and text data acquired from the flower dataset. It entails converting images to NumPy arrays, modifying their pixel sizes to preset criteria, and storing these modified arrays for later model input. Concurrently, the textual descriptions that accompany the images are converted into embeddings and saved in a structured CSV file format. The incorporation of these preprocessed data pieces is critical, as it serves as the fundamental input for the future stages and ensures compliance with the GAN architecture, which is meant to convert textual descriptions into realistic visual outputs.
4. **Modelling:** During the modelling phase, the focus is on constructing a unique deep architecture comprised of the generator and discriminator networks by the CRISP-DM framework for text-to-image creation utilising GANs. This entails building these neural network components, specifying their complicated architecture, and implementing methods to calculate discriminator and generator losses.

Setting the optimizer and learning rates is critical for improving model performance. This phase focuses on designing the GAN and developing the key components required for translating textual descriptions into visually realistic images. This model's rigorous design and development lay the basis for further training and assessment stages, to produce high-quality images from comprehensive text inputs.

5. **Evaluation:** In the evaluation phase following the CRISP-DM framework for text-to-image generation using GANs, the model's performance is rigorously assessed and analysed. This involves calculating the loss and adjusting the gradients. This phase aims to comprehensively evaluate the effectiveness and reliability of the text-to-image generation model, shedding light on its suitability for real-world applications and guiding potential improvements or future iterations.

3.2 Model Training

In the model training phase, crucial for text-to-image generation using GANs, specific functions are implemented to facilitate the iterative improvement of the GAN model. The 'train_step' function plays a pivotal role in this process, creating images based on textual descriptions, calculating the associated losses (typically including adversarial and reconstruction losses), and adjusting gradients to enhance the model's performance. Additionally, the overarching 'train' function orchestrates the entire training process by fetching batch data, passing it through the 'train_step' function, and collecting the resultant losses. This iterative training mechanism refines the GAN architecture's capabilities, iteratively optimizing it to generate more realistic and visually compelling images from textual inputs. The coordinated interplay between these functions is fundamental in enhancing the model's ability to effectively translate textual descriptions into plausible and high-quality image outputs.

3.3 Model Building

In the process of model building for text-to-image generation using GANs, two fundamental components are developed: the Generator Layer and the Discriminator Layer. The Generator Layer is responsible for creating images based on textual descriptions. It comprises a neural network structure designed to transform input text or latent vectors into visually realistic images. Conversely, the Discriminator Layer is tasked with discerning between real images from the dataset and those generated by the Generator Layer.

Figure : Generator Layer

Figure : Discriminator Layer

3.4 Dataset Description

The dataset utilized in this project comprises 102 distinct categories of flower images, meticulously curated to represent commonly occurring flowers in the United Kingdom. Each category contains a variable number of images, ranging from 40 to 258 images per class. This diversity in image quantity across categories enriches the dataset, offering a comprehensive representation of various flower species. Each image is meticulously annotated with descriptive sentences, providing context and information about the depicted flowers. Notably,

the dataset exhibits extensive variability in terms of scale, pose, and lighting conditions, capturing the inherent complexities and nuances present in real-world flower photography. Additionally, certain categories exhibit significant intra-category variations, contributing to the dataset's richness, while other categories showcase remarkable similarities. This diversity and intricacy are visualized using isomap, revealing the dataset's complexity through shape and colour features, emphasizing the challenges and opportunities for image recognition and generation tasks. The dataset's unique characteristics, encompassing variations within and between categories, render it an ideal resource for training and evaluating models aimed at tasks like image classification, generation, and semantic understanding of floral imagery.

Chapter 4 Design Specification and Implementation

The design specification chapter encapsulates the architectural blueprint and technical intricacies of the text-to-image generation system utilizing Generative Adversarial Networks (GANs). At its core, the system revolves around a novel deep architecture merging the power of recurrent neural networks and convolutional GANs. The primary goal is to bridge the gap between textual descriptions and visually plausible image outputs. The architectural design is segmented into key components: the Generator Layer, tasked with transforming text descriptions into high-resolution images, and the Discriminator Layer, responsible for discerning between real images from the dataset and those generated by the Generator. The Generator Layer, a pivotal element, harnesses recurrent networks or transposed convolutions to interpret text embeddings and create corresponding image representations. In tandem, the Discriminator Layer employs a separate neural network to scrutinize the generated images, providing feedback to enhance the Generator's ability to produce more authentic outputs. This adversarial training mechanism facilitates the iterative refinement of both the Generator and Discriminator networks, fostering an environment where they continuously improve their abilities through competition.

Within this design, careful consideration is given to the preprocessing pipeline, ensuring seamless integration of image and text data. Preprocessing involves converting images into standardized NumPy arrays, adjusting pixel sizes as per predetermined specifications, and transforming textual descriptions into embeddings stored in a structured CSV format. The system architecture also encompasses functions crucial for model training, such as the 'train_step' function, responsible for creating images based on text descriptions, calculating losses,

and adjusting gradients, and the overarching 'train' function, orchestrating the training process by fetching batch data and aggregating losses.

Moreover, the system implements optimization techniques like specific loss functions, optimizers, and learning rates, critical in refining the GAN model's performance. Evaluation metrics, including image similarity scores and subjective assessments, will gauge the model's efficacy in generating realistic images from textual inputs. To ensure scalability and usability, the system is designed to be adaptable, with potential applications across various domains such as e-commerce, design, and content creation. This comprehensive design specification lays the groundwork for the development, evaluation, and potential deployment of a robust text-to-image generation system.

Chapter 5 Evaluation

The Image Generator Model's outputs represent a great achievement after a lengthy training period of roughly 450 epochs and a total training duration of 24 hours. The design of the model included both the Discriminator and Generator networks, which are critical components of the GAN framework for text-to-image creation. Specific functions were used to calculate losses for both the Discriminator and Generator networks during the training phase. Therefore, visible progress was made by the 438th epoch, with the Generator loss measured at 1.3284586668014526 and the Discriminator loss measured at 1.2313429117202759. The achieved losses imply that the Generator and Discriminator networks are approaching equilibrium. A decreasing Generator loss implies that the model's capacity to create more realistic images from textual descriptions is improving. Meanwhile, the Discriminator loss, while greater than the Generator loss as predicted in adversarial training, represents the network's capacity to distinguish between actual and created images.

5.1 Performance Evaluation on Test Data

Figure 5.1 Purple Flower with Oval-Shaped Petals

In Figure 5.1 of the report, the showcased test_image portrays a flower characterized by a purple hue, exhibiting oval-shaped petals. The textual description accompanying this image emphasizes these distinct visual attributes. Subsequently, upon examining the test_output result, it reveals a multitude of smaller images. These images are likely the generated outputs produced by the model in response to the provided textual description. The presence of numerous smaller images suggests the model's attempts to interpret and generate various visual representations corresponding to the textual description of a purple-coloured flower with oval-shaped petals. This observation underscores the model's efforts in generating diverse outputs to encapsulate the potential variations and nuances inherent in the given textual input, aiming to produce a range of plausible floral images that align with the described characteristics. Further evaluation and analysis of these outputs would provide deeper insights into the model's ability to interpret textual descriptions and generate corresponding diverse and realistic visual representations.

Figure 5.2 Yellow Flower with Oval-Shaped Petals

Figure 5.2 illustrates a test_image depicting a flower described as yellow with oval-shaped petals. This textual description accompanies the displayed image, emphasizing its distinct visual attributes. The test_output result presented in Figure 5.2 showcases a collection of smaller images. These smaller images likely represent the model's generated outputs corresponding to the provided textual description. The abundance of diverse smaller images indicates the model's attempts to interpret and generate various visual representations in response to the description of a yellow-coloured flower with oval-shaped petals. This multitude of outputs suggests the model's endeavour to capture the potential variations and intricacies inherent in the textual input, aiming to produce a range of plausible floral images that align with the specified characteristics. Further analysis and assessment of these generated outputs would provide deeper insights into the model's capacity to comprehend textual descriptions and generate diverse and realistic visual depictions.

The extensive training duration of 24 hours over nearly 450 epochs underline the complexity and computational intensity of the model. This prolonged training period was crucial for the convergence of the networks, ensuring incremental improvements in image generation quality over time. Despite these accomplishments, this phase represents a major checkpoint rather than a definitive ending. Beyond epoch 450, more study is required to determine the model's continuous development and convergence. Furthermore, fine-tuning factors such as optimizer configurations and learning rates may lead to further performance improvements. While the achieved losses indicate promising progress, the actual quality and realism of the produced images need rigorous examination utilising quantitative measures and subjective judgments.

Moving forward, this stage sets the foundation for a comprehensive evaluation, where both quantitative metrics and qualitative judgments by human assessors will scrutinize the generated images' fidelity and

realism. These assessments will provide deeper insights into the model's effectiveness, refining the system and determining its readiness for practical applications.

5.2 Text to Image Generator (GUI Result)

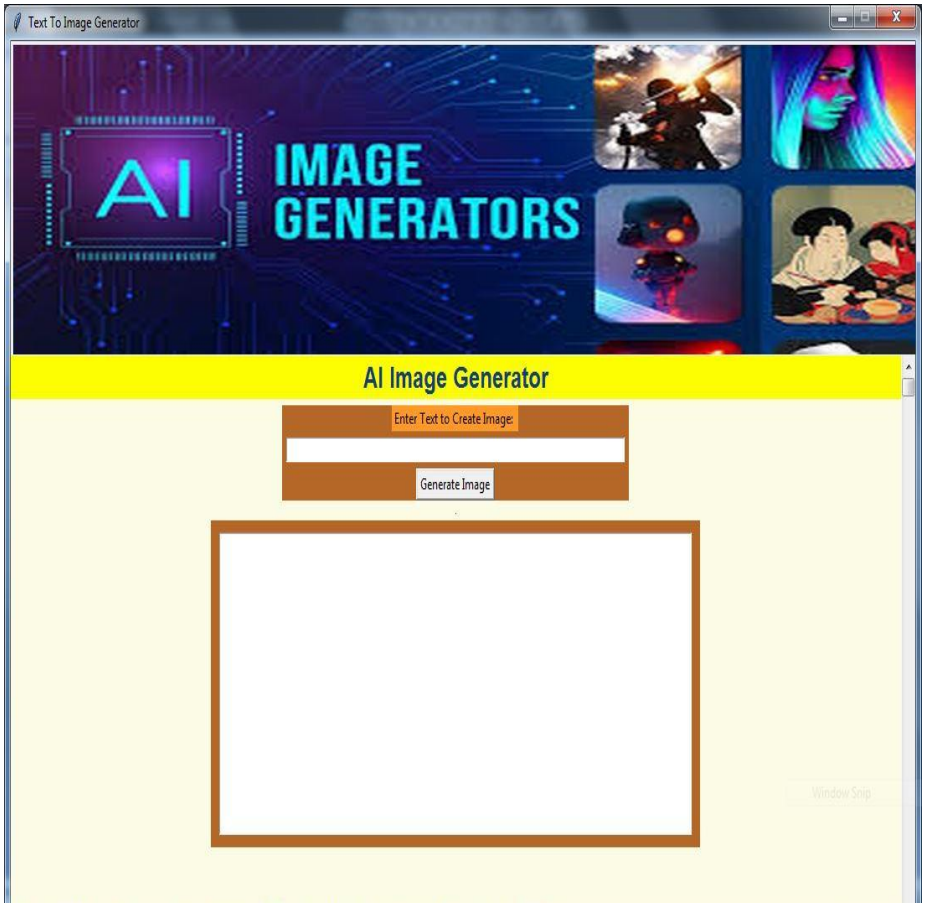

Figure 5.3: AI-Powered Text-to-Image Generator Interface

Figure 5.3 showcases an AI-powered text-to-image generator interface. The displayed image represents the user interface or platform designed for generating images from textual descriptions using artificial intelligence. The interface features a message prompting the user: "Enter text to create Image then click on generate image." This message serves as an instruction or guidance for the user, outlining the required steps to utilise the image generation functionality offered by the AI system. Users are encouraged to input textual descriptions of the desired image content into the provided text entry field. Once the desired text is entered, the user initiates the image generation process by clicking on the "generate image" button or similar interface element. This action prompts the AI image generator to interpret the input text and produce corresponding visual outputs, generating images that encapsulate the semantic content described in the entered text. The interface design aims to facilitate user interaction, enabling individuals to effortlessly create images based on

their textual descriptions through a straightforward and intuitive process within the AI-powered image generation platform.

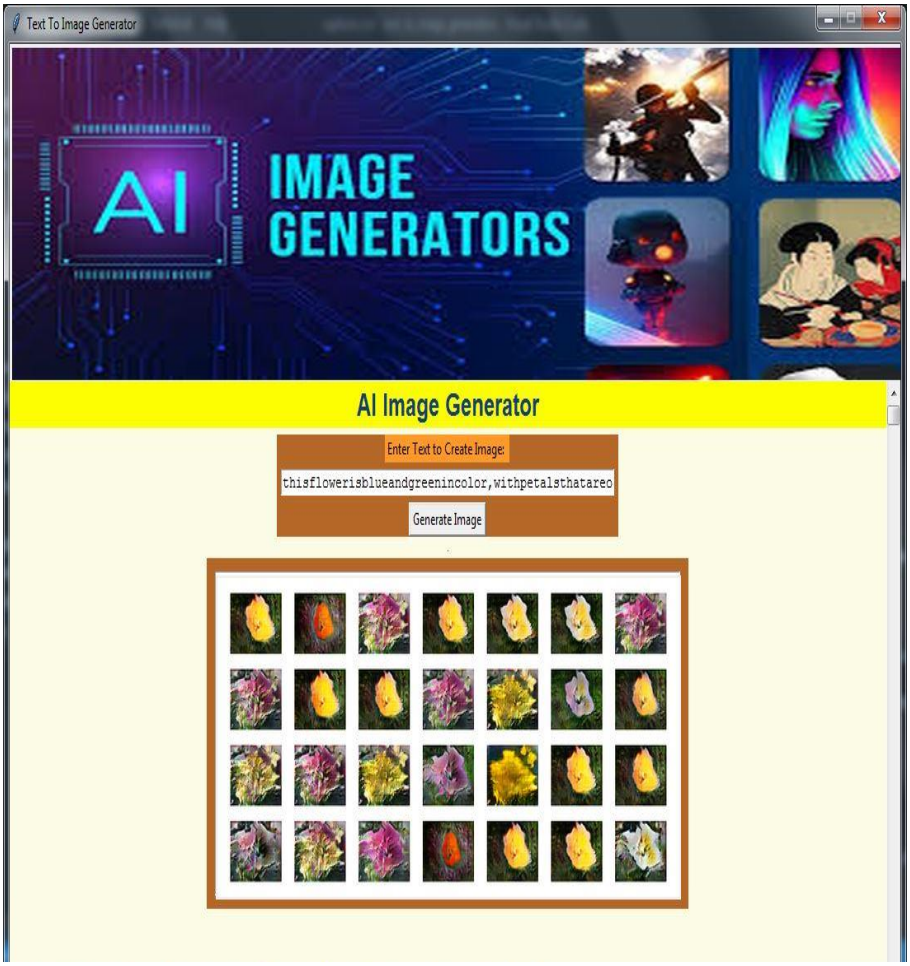

Figure 5.4: AI-Driven Text-to-Image Generation Interface: Flower with Oval-Shaped Petals

Figure 5.4 presents an interface that facilitates the generation of AI-created images using a text-to-image generator system. The displayed interface gives users a text input field and a description: "This will generate a Flower with Oval-Shaped Petals." This text serves as an example or instruction guiding users on the type of input to provide for image generation. By entering similar descriptive text into the designated field, such as "Flower with Oval-Shaped Petals," users can prompt the AI-powered system to interpret the textual description and generate corresponding visual outputs—specifically, images that represent the described concept. This interface showcases the system's capability to translate textual information into visual content, enabling users to generate images based on their provided descriptions through an intuitive and user-friendly interface. The text input serves as a directive for users, illustrating the type of input expected to prompt the generation of specific image content by the AI text-to-image generator system displayed in Figure 5.4.

Chapter 6 Conclusion and Future Works

6.1 Conclusions

Significant milestones and discoveries have occurred throughout the development of a text-to-image generating model employing GANs. Convergence in the Generator and Discriminator networks, as indicated by lowering Generator loss and maintaining Discriminator loss, demonstrates potential progress in

synthesising realistic images from textual descriptions. The lengthy training process, which lasted around 450 epochs over 24 hours, reflects the model's intricacy and the diligent efforts made to optimise its capabilities. This checkpoint is a critical stage in the development of the model, establishing the framework for detailed assessment and potential practical implementations.

6.2 Limitations

Despite development, several limits remain. The present assessment, which focuses on losses and convergence, does not fully reflect the quality and perceptual integrity of produced images. To support quantitative measurements, subjective assessments and sophisticated qualitative evaluations are required. Furthermore, the computational intensity and time required for training need optimization measures to improve efficiency and scalability.

6.3 Future Works

Future enhancements and research avenues encompass multifaceted dimensions. Comprehensive evaluation involving human raters' subjective assessments will gauge the perceptual fidelity and realism of generated images, providing nuanced insights. Fine-tuning parameters, such as optimizing optimizer configurations and fine-tuning learning rates, can further elevate the model's performance. Exploration of advanced GAN variants or attention mechanisms might enrich the model's ability to capture intricate details and improve image quality. Furthermore, considerations for scalability and efficiency, possibly through parallel computing or hardware acceleration, will expedite model training and deployment.

In conclusion, while this phase signifies substantial progress in text-to-image generation, a comprehensive evaluation, fine-tuning, and exploration of advanced techniques remain crucial for refining the model's capabilities and ensuring its practical applicability across diverse domains. This checkpoint sets the trajectory for continued research, aiming to push the boundaries of text-to-image generation technology and unlock its potential in various real-world applications.

References

1. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M. and Sutskever, I., 2021, July. Zero-shot text-to-image generation. In *International Conference on Machine Learning* (pp. 8821-8831). PMLR.
2. Sharma, S., Suhubdy, D., Michalski, V., Kahou, S.E. and Bengio, Y., 2018. Chatpainter: Improving text to image generation using dialogue. *arXiv preprint arXiv:1802.08216*.
3. Dong, H., Zhang, J., McIlwraith, D. and Guo, Y., 2017, September. I2t2i: Learning text to image synthesis with textual data augmentation. In *2017 IEEE international conference on image processing (ICIP)* (pp. 2015-2019). IEEE.
4. Qiao, T., Zhang, J., Xu, D. and Tao, D., 2019. Mirrorgan: Learning text-to-image generation by redescription. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1505-1514).

5. Zia, T., Arif, S., Murtaza, S. and Ullah, M.A., 2020. Text-to-image generation with attention based recurrent neural networks. *arXiv preprint arXiv:2001.06658*.
6. He, X. and Deng, L., 2017. Deep learning for image-to-text generation: A technical overview. *IEEE Signal Processing Magazine*, *34*(6), pp.109-116.
7. Chen, J. and Zhuge, H., 2018, September. Extractive text-image summarization using multi-modal RNN. In *2018 14th International Conference on Semantics, Knowledge and Grids (SKG)* (pp. 245-248). IEEE.
8. Chen, J. and Zhuge, H., 2018. Abstractive text-image summarization using multi-modal attentional hierarchical rnn. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 4046-4056).
9. Ramzan, S., Iqbal, M.M. and Kalsum, T., 2022. Text-to-Image Generation Using Deep Learning. *Engineering Proceedings*, *20*(1), p.16.
10. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B. and Lee, H., 2016, June. Generative adversarial text to image synthesis. In *International conference on machine learning* (pp. 1060-1069). PMLR.
11. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X. and He, X., 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1316-1324).
12. Liao, W., Hu, K., Yang, M.Y. and Rosenhahn, B., 2022. Text to image generation with semantic-spatial aware gan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 18187-18196).
13. Zhu, M., Pan, P., Chen, W. and Yang, Y., 2019. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5802-5810).
14. Liu, X., Gong, C., Wu, L., Zhang, S., Su, H. and Liu, Q., 2021. Fusedream: Training-free text-to-image generation with improved clip+ gan space optimization. *arXiv preprint arXiv:2112.01573*.
15. Ruan, S., Zhang, Y., Zhang, K., Fan, Y., Tang, F., Liu, Q. and Chen, E., 2021. Dae-gan: Dynamic aspect-aware gan for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 13960-13969).
16. Wang, H., Lin, G., Hoi, S.C. and Miao, C., 2021, October. Cycle-consistent inverse GAN for text-to-image synthesis. In *Proceedings of the 29th ACM International Conference on Multimedia* (pp. 630-638).
17. Zhang, C. and Peng, Y., 2018, September. Stacking VAE and GAN for context-aware text-to-image generation. In *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)* (pp. 1-5). IEEE.
18. Tan, H., Liu, X., Yin, B. and Li, X., 2022. DR-GAN: Distribution regularization for text-to-image generation. *IEEE Transactions on Neural Networks and Learning Systems*.
19. Sawant, R., Shaikh, A., Sabat, S. and Bhole, V., 2021, July. Text to image generation using GAN. In *Proceedings of the International Conference on IoT Based Control Networks & Intelligent Systems-ICICNIS*.
20. Tao, M., Tang, H., Wu, F., Jing, X.Y., Bao, B.K. and Xu, C., 2022. Df-gan: A simple and effective baseline for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 16515-16525).
21. Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H. and Tang, J., 2021. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, *34*, pp.19822-19835.

22. Khan, W., Daud, A., Khan, K., Muhammad, S. and Haq, R., 2023. Exploring the frontiers of deep learning and natural language processing: A comprehensive overview of key challenges and emerging trends. *Natural Language Processing Journal*, p.100026.

23. Huang, M., Mao, Z., Wang, P., Wang, Q. and Zhang, Y., 2022, October. Dse-gan: Dynamic semantic evolution generative adversarial network for text-to-image generation. In *Proceedings of the 30th ACM International Conference on Multimedia* (pp. 4345-4354).

24. Zhou, Y., Zhang, R., Chen, C., Li, C., Tensmeyer, C., Yu, T., Gu, J., Xu, J. and Sun, T., 2022. Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 17907-17917).

25. Ganz, R. and Elad, M., 2024. Clipag: Towards generator-free text-to-image generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 3843-3853).

26. Berrahal, M. and Azizi, M., 2022. Optimal text-to-image synthesis model for generating portrait images using generative adversarial network techniques. *Indones. J. Electr. Eng. Comput. Sci*, *25*(2), pp.972-979.