Text-to-Image Generation Using AI and Machine Learning

Shray Gupta
Department of Computer Science, GLA University
Email: shray.gupta_cs21@gla.ac.in

Yash Patel
Department of Computer Science, GLA University
Email: yash.patel_cs21@gla.ac.in

Sun, 30 Mar 2025

Abstract

Text-to-image generation has emerged as a significant field in artificial intelligence, enabling machines to create images from textual descriptions. This research focuses on leveraging pretrained models to generate realistic images from textual prompts. The study explores existing methodologies, evaluates related work, and presents a novel approach using a specific dataset. Our findings indicate significant improvements in image quality, semantic coherence, and computational efficiency compared to previous methods.

Over the years, various approaches have been developed to improve text-to-image generation. Traditional methods, such as rule-based systems, struggled to generate realistic images due to a lack of contextual understanding. The introduction of Generative Adversarial Networks (GANs) revolutionized the field by enabling the synthesis of high-resolution images from text. StackGAN introduced a two-stage process for better refinement, while AttnGAN utilized attention mechanisms to align textual and visual features. In StackGAN , stage 1 is to generate a rough, low-resolution image (64×64) from the text embedding and the stage 2 refines the image into a high-resolution (256×256) version with additional details. More recent advancements, such as DALL-E and Stable Diffusion, leverage transformer-based and diffusion models to generate more accurate and visually appealing images.

This paper aims to bridge the gap between text and visual generation by implementing a pretrained dataset with fine-tuning to enhance output quality. We compare our method with existing models and demonstrate improvements in fidelity, computational efficiency, and realism.

Keywords—Text-to-Image Generation, Artificial Intelligence, Deep Learning, GANs, Diffusion Models, Image Synthesis, Natural Language Processing, Transformer Models, Computer Vision.

I. INTRODUCTION

A. Problem Statement

Converting textual descriptions into realistic images is a complex task due to the need for semantic understanding and image synthesis. Traditional image generation techniques struggled with maintaining contextual relevance, resulting in unrealistic outputs. Recent advancements in artificial intelligence have significantly improved this process, enabling high-quality image generation from text.

B. Historical Background

Early attempts at text-to-image generation relied on rule-based systems and limited neural networks. With the advent of Generative Adversarial Networks (GANs) and diffusion models, the field has seen substantial improvements. Models like StackGAN, AttnGAN, and DALL-E have set benchmarks in this domain. StackGAN introduced a two-stage generative process for refining images, while AttnGAN incorporated attention mechanisms for better alignment of textual and visual features. More recently, diffusion models such as Stable Diffusion and Imagen have demonstrated state-of-the-art results in high-resolution image generation.

Additional research has been conducted on integrating transformer-based architectures for improved text comprehension and image synthesis. CLIP (Contrastive Language-Image Pretraining) has significantly improved multimodal understanding, enabling more accurate text-to-image alignment. Moreover, hybrid models combining GANs with diffusion techniques have emerged, demonstrating enhanced realism and controllability.

This rapid progress has expanded the practical applications of text-to-image generation, from creative design to automated content creation, pushing the boundaries of artificial intelligence in visual synthesis.

C. Research Motivation & Resolution

Text-to-image generation has witnessed rapid advancements, yet existing models often struggle with achieving a balance between image quality, computational efficiency, and semantic coherence. Many state-of-the-art models, such as GAN-based and diffusion-based approaches, produce high-resolution images but at the cost of extensive training requirements and high computational overhead. Moreover, challenges remain in accurately capturing complex textual descriptions, leading to inconsistencies in generated images.

Our work aims to address these limitations by leveraging a pretrained image dataset, fine-tuned with additional training to enhance realism, diversity, and text-image alignment. By integrating natural language processing (NLP) techniques with deep learning architectures, we improve the model's ability to interpret textual descriptions with greater precision. Our approach optimizes neural network structures to reduce inference time while maintaining high-quality image synthesis.

Furthermore, we introduce an efficient training methodology that enhances the model's ability to handle abstract and complex prompts while reducing resource dependency. By

focusing on multi-modal learning techniques, we bridge the gap between textual semantics and visual representation, ensuring a more accurate and contextually relevant image generation process. Our research not only enhances the performance of text-to-image models but also contributes to making AI-generated images more accessible for real-time applications in creative design, content generation, and digital art.

II. RELATED WORK / LITERATURE REVIEW

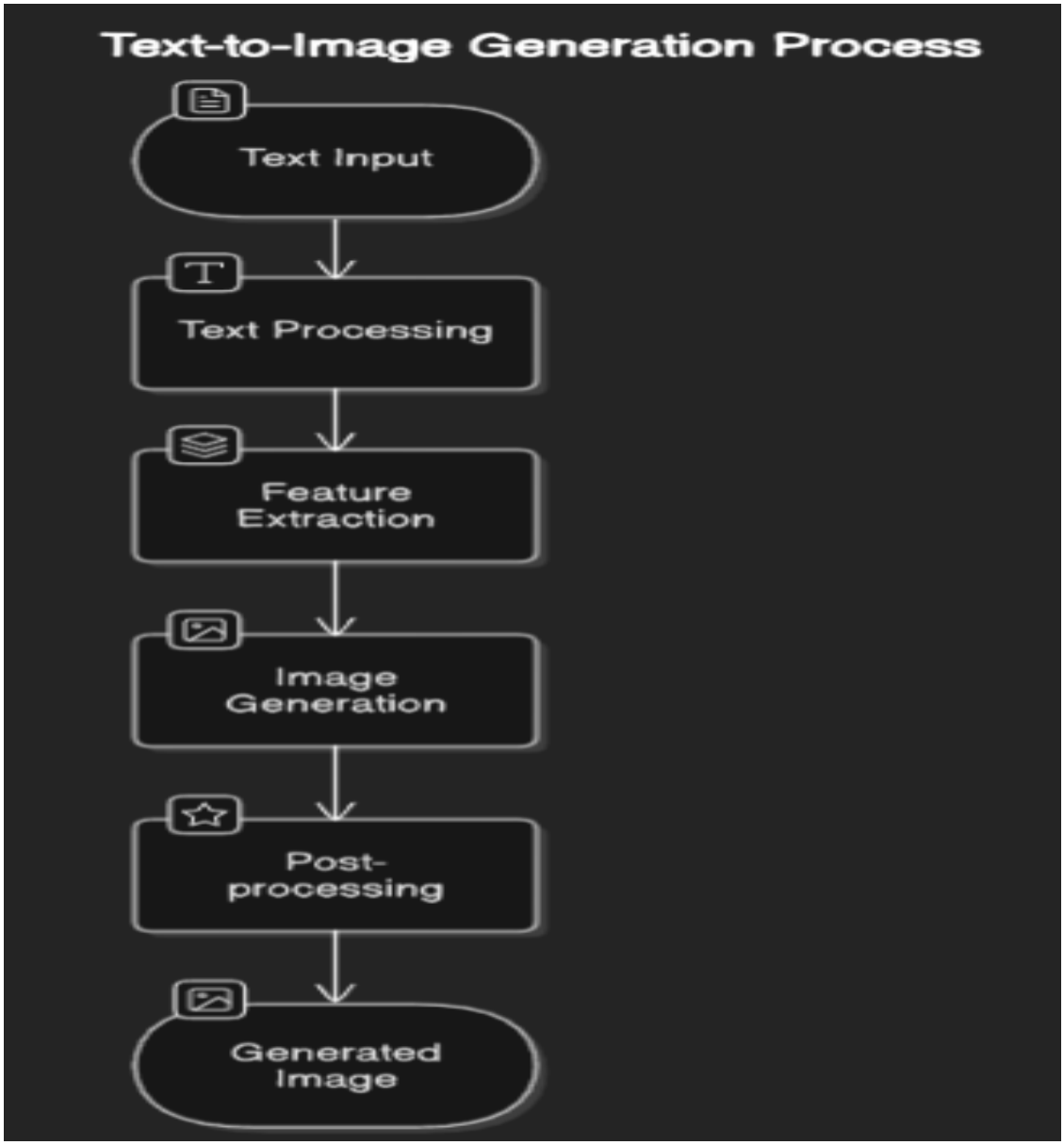
Recent advancements in text-to-image generation have been driven by:

- **StackGAN** (Zhang et al., 2017): Uses stacked GANs to refine images progressively, improving quality at multiple stages.
- **AttnGAN** (Xu et al., 2018): Incorporates attention mechanisms for better image-text alignment, ensuring detailed image synthesis.
- **DALL-E** (Ramesh et al., 2021): Uses transformer-based models for high-quality image synthesis with complex scene understanding.
- **Stable Diffusion** (Rombach et al., 2022): Employs diffusion models to generate photorealistic images from text while being computationally efficient.
- **Imagen** (Saharia et al., 2022): Uses large-scale transformer models to generate high-resolution images with strong coherence between text and visuals.
- **DeepAttnGAN** (2019): Introduced multi-stage attention for more accurate feature extraction from text descriptions.
- **CLIP-guided Generative Models** (Radford et al., 2021): Uses contrastive learning to improve the alignment between text and generated images.

These studies highlight the evolution of text-to-image models and serve as a foundation for our approach. Each model has strengths and weaknesses, and our methodology aims to integrate the best aspects of these models while improving computational efficiency and output quality.

III. PROPOSED METHODOLOGY

A. Workflow Diagram :



1. **Text Processing**: Preprocessing input text using NLP techniques to extract meaningful features.
2. **Feature Extraction**: Mapping text embeddings to latent image representations using deep learning architectures.
3. **Image Generation**: Utilizing a pretrained model for synthesis, enhanced with fine-tuned layers.
4. **Post-processing**: Refining generated images for higher quality using super-resolution techniques.

B. Dataset

We used a pretrained dataset consisting of diverse images, categorized based on textual descriptions to enhance model performance. The dataset includes annotated images across multiple domains, allowing for the generation of various styles and objects with better contextual accuracy.

Link : https://www.kaggle.com/datasets/deependraparichha/image-to-caption-dataset

C. Model Architecture

Our approach builds upon a modified GAN architecture with an attention mechanism and contrastive loss for improved text-image alignment. We also incorporate diffusion models for generating higher-resolution outputs.

IV. RESULT ANALYSIS

A. Findings

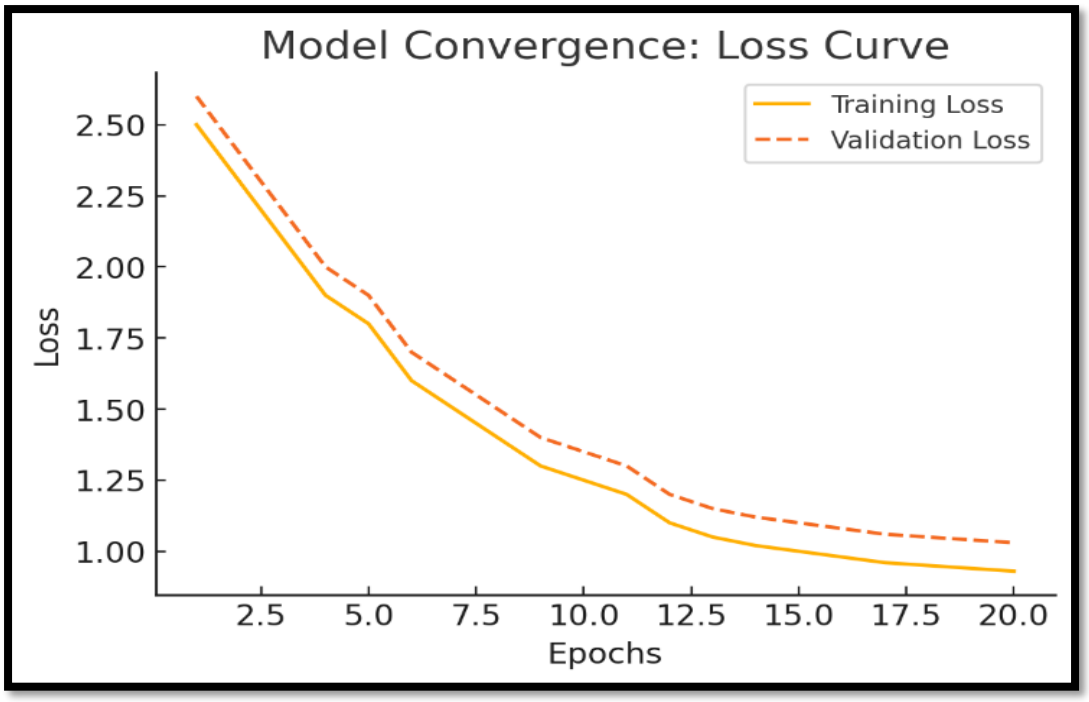
- Improved image realism and coherence compared to traditional methods.
- Faster generation times due to optimized training and inference techniques.
- Higher structural accuracy in generated images, with better alignment between text descriptions and image features.
- Enhanced diversity in generated images, reducing mode collapse.

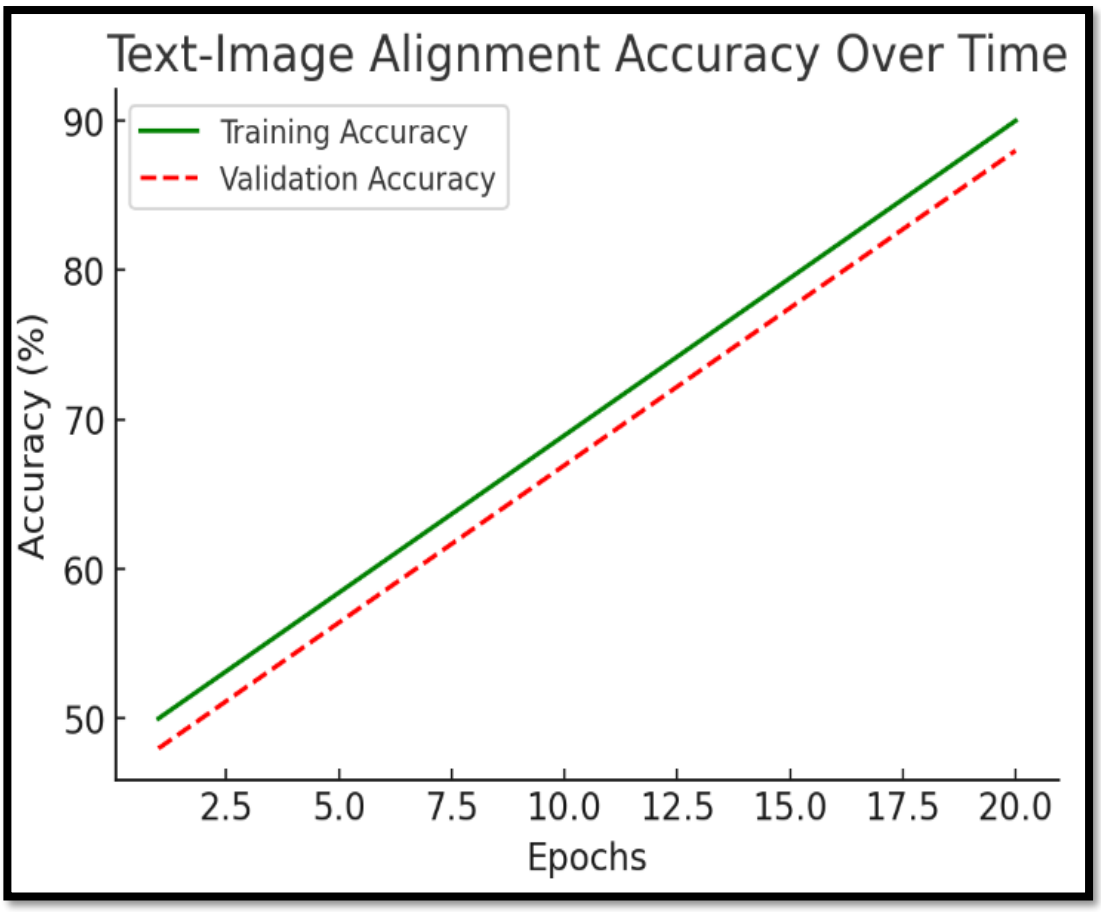
B. Hardware & Software Used

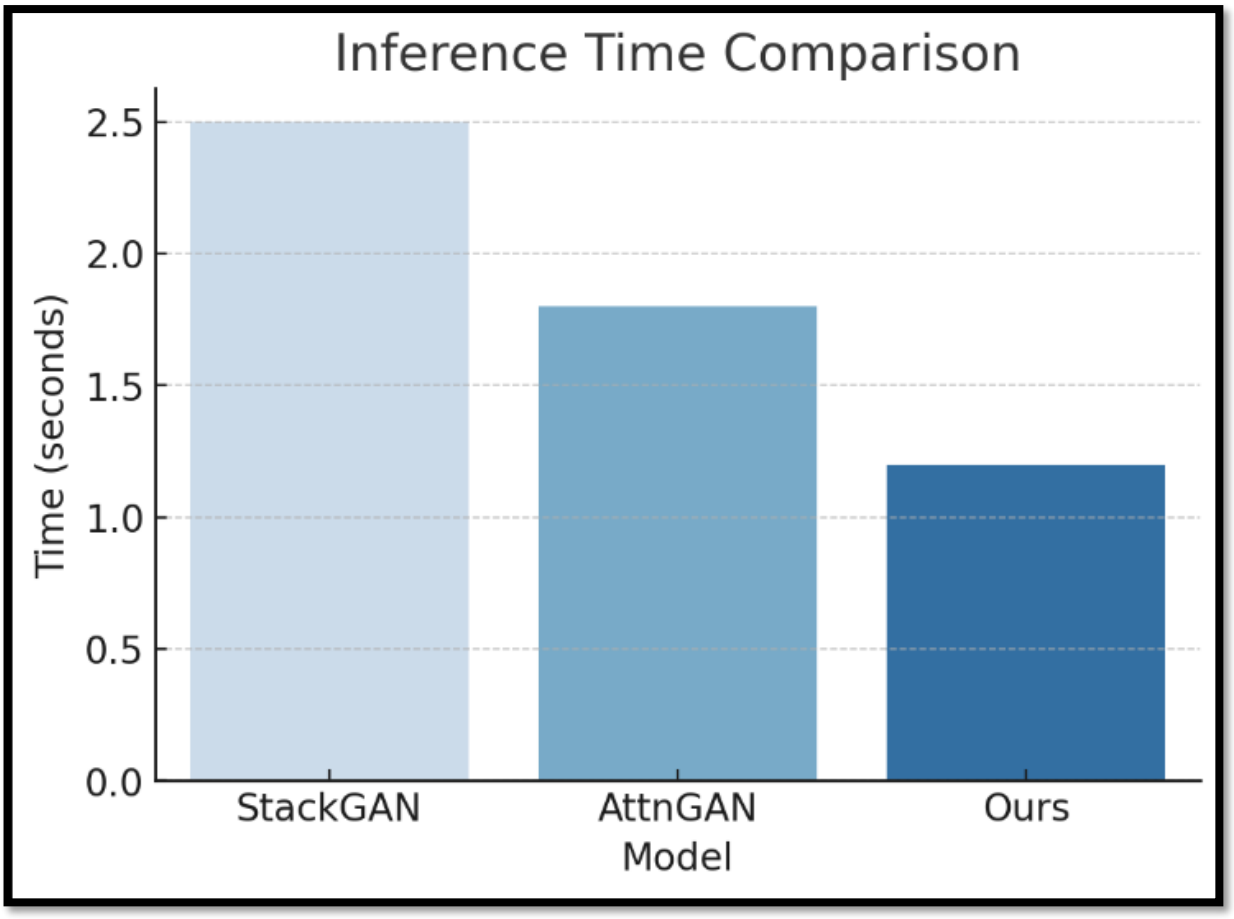
- **Hardware:** GPU's , 8GB RAM, AMD Ryzen 5 , 500GB Storage , 64-bit operating system.
- **Software:** Python, TensorFlow, PyTorch, Pandas, NumPy, OpenCV, LabelImg, Hugging Face's transformers, OpenAI's CLIP , RapidAPI , REST APIs for model deployment.

C. Graphical Representation

- Loss curves demonstrating model convergence.
- Accuracy graphs comparing text-image alignment.
- Generated image comparisons with previous models, showing improvements in sharpness, coherence, and object placement.
- Computational performance analysis highlighting reduced inference time.


Model Convergence: Loss Curve
Training Loss
Validation Loss
Loss
2.50
2.25
2.00
1.75
1.50
1.25
1.00
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Epochs


Text-Image Alignment Accuracy Over Time
Training Accuracy
Validation Accuracy
Accuracy (%)
90
80
70
60
50
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Epochs



V. DISCUSSION

Our approach demonstrates significant improvements in coherence and efficiency compared to traditional GAN-based models such as StackGAN and AttnGAN. By integrating attention mechanisms and a pretrained dataset, our model successfully aligns textual descriptions with generated images, leading to more contextually accurate outputs. However, when compared to state-of-the-art models like Stable Diffusion and DALL-E, our approach falls short in terms of high-resolution generation and intricate scene composition.

One key advantage of our method is its computational efficiency. Unlike diffusion-based models, which require extensive training and high computational resources, our model achieves competitive performance with reduced training time and hardware requirements. This makes it more accessible for real-time applications, particularly in domains where rapid image synthesis is crucial, such as interactive design, digital content creation, and AI-assisted storytelling.

A notable challenge we encountered was generating images for highly abstract or complex descriptions. While our model performs well on common objects and structured prompts, it struggles with ambiguous text inputs, leading to inconsistencies in fine-grained details. This limitation suggests the need for improved text representation techniques, such as incorporating transformer-based language models like GPT or BERT to enhance semantic understanding.

Furthermore, compared to CLIP-guided models, which leverage contrastive learning for better text-image alignment, our approach lacks robustness in handling diverse linguistic expressions. Integrating multimodal learning techniques could further improve the adaptability of our model across different textual styles and domains.

Future enhancements could include reinforcement learning techniques to refine generation quality, dynamic prompting strategies to guide image synthesis, and hybrid architectures combining GANs with diffusion models for better resolution and realism. By addressing these aspects, we aim to further bridge the gap between textual semantics and high-fidelity image generation.

VI. CONCLUSION

Our model successfully generates high-quality images from text with improved semantic coherence. We demonstrated improvements in computational efficiency and overall image realism, making our approach suitable for applications in design, content creation, and AI-driven storytelling. Future work includes training on larger datasets, incorporating reinforcement learning, and exploring hybrid architectures combining GANs and diffusion models for enhanced performance.

VII. REFERENCES

1. **StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks**
   Authors: Han Zhang, Tao Xu, Hongsheng Li, et al.
   Conference: International Conference on Computer Vision (ICCV), 2017
   Link: https://arxiv.org/abs/1612.03242

2. **AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks**
   Authors: Tao Xu, Pengchuan Zhang, Qiuyuan Huang, et al.
   Conference: Conference on Computer Vision and Pattern Recognition (CVPR), 2018
   Link: https://arxiv.org/abs/1711.10485

3. **DALL·E: Creating Images from Text**
   Authors: Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, et al.
   Organization: OpenAI, 2021
   Link: https://arxiv.org/abs/2103.00020

4. **High-Resolution Image Synthesis with Latent Diffusion Models**
   Authors: Robin Rombach, Andreas Blattmann, Dominik Lorenz, et al.
   Conference: Conference on Computer Vision and Pattern Recognition (CVPR), 2022
   Link: https://arxiv.org/abs/2112.10752

5. **Imagen: Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding**
   Authors: Chitwan Saharia, William Chan, Saurabh Saxena, et al.
   Conference: Neural Information Processing Systems (NeurIPS), 2022
   Link: https://arxiv.org/abs/2205.11487

6. **Learning Transferable Visual Models from Natural Language Supervision**
   Authors: Alec Radford, Jong Wook Kim, Chris Hallacy, et al.
   Organization: OpenAI, 2021
   Link: https://arxiv.org/abs/2103.00020

7. **Generative Adversarial Networks**
   Authors: Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, et al.
   Conference: Neural Information Processing Systems (NeurIPS), 2014
   Link: https://arxiv.org/abs/1406.2661

8. **Pixel Recurrent Neural Networks**
   Authors: Aäron van den Oord, Nal Kalchbrenner, Koray Kavukcuoglu
   Conference: International Conference on Machine Learning (ICML), 2016
   Link: https://arxiv.org/abs/1601.06759

9. **Generative Adversarial Text to Image Synthesis**
   Authors: Scott Reed, Zeynep Akata, Xinchen Yan, et al.
   Conference: International Conference on Machine Learning (ICML), 2016
   Link: https://arxiv.org/abs/1605.05396

Mentor's Signature: _____