International Journal of Scientific Research and Engineering Development-– Volume 8 Issue 3, May-June 2025
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

A Review on AI-Driven Text-to-Image Generation Using Deep Learning and GANs

[1]Ms.Aachal G. Gajbhiye,[2]Dr. R. R. Keole,[3]Dr.A. P. Jadhao,[4]Prof. D. G. Ingale
1. ME Student, Dr. Rajendra Gode Institute of Technology & Research, Amravati
2. Guide, HVPM College of Engineering and Technology, Amravati
3. Co-Guide, Dr. Rajendra Gode Institute of Technology & Research, Amravati
4. ME Coordinator, Dr. Rajendra Gode Institute of Technology & Research, Amravati

Abstract
Text-to-image generation is a rapidly evolving field in artificial intelligence that focuses on synthesizing realistic images from natural language descriptions. This process lies at the intersection of computer vision and natural language processing, requiring a deep semantic understanding of both modalities. The introduction of Generative Adversarial Networks (GANs) has significantly enhanced the quality, resolution, and diversity of generated images. Models such as StackGAN, AttnGAN, and DM-GAN have demonstrated notable advancements by incorporating multi-stage architectures, attention mechanisms, and dynamic memory modules to improve the semantic alignment between text and images [9], [6], [2]. Recent developments have also scaled up GAN architectures to produce more photorealistic and controllable outputs [2], [3]. This review paper provides a comprehensive analysis of AI-driven techniques for text-to-image generation, emphasizing deep learning frameworks and GAN-based methods. It discusses model architectures, benchmark datasets, evaluation metrics, challenges, and future research directions to guide further progress in this transformative area of research.

I. Introduction

The task of transforming textual descriptions into visual representations, known as text-to-image generation, has become a focal point of research at the intersection of computer vision and natural language processing. This interdisciplinary challenge requires models that can comprehend and bridge the semantic gap between language and imagery. In recent years, artificial intelligence (AI), particularly deep learning, has played a pivotal role in addressing this task by enabling systems to learn complex mappings between text and images in a data-driven manner.

The advent of Generative Adversarial Networks (GANs) [1] revolutionized generative modeling by introducing a competitive training framework between a generator and a discriminator. This architecture laid the groundwork for conditional GANs (cGANs), which enabled the generation of images based on auxiliary information such as class labels or text descriptions [2]. Since then, numerous models have emerged, aiming to improve image fidelity, semantic alignment, and text-image coherence.

One of the early breakthroughs in this domain was the StackGAN model, which used a multi-stage generative process to produce high-resolution images from textual input [9]. This was followed by architectures like AttnGAN and DM-GAN, which incorporated attention mechanisms and dynamic memory modules to enhance the interpretability and quality of generated images [6], [2]. More recently, scalable models such as StyleGAN-T and other large-scale GAN-based frameworks have further advanced the realism and controllability of generated outputs [3], [2].

Despite these advancements, text-to-image generation still faces several challenges, including the need for large annotated datasets, training instability, and semantic mismatches between the textual prompt and the visual output. To address these issues, researchers have explored improvements in model architecture, loss functions, and evaluation metrics [4], [7].

This review aims to provide a comprehensive overview of the current state of AI-driven text-to-

 ©IJSRED: All Rights are Reserved

image generation. It highlights the evolution of deep learning-based approaches, focusing on GANs, explores widely used datasets and evaluation techniques, and discusses the key challenges and future opportunities in this fast-growing field.

II. Background and Fundamentals

2.1 Deep Learning Basics

Deep learning, a subfield of machine learning, has gained prominence due to its ability to automatically extract relevant features from large amounts of unstructured data. Neural networks, particularly **Convolutional Neural Networks (CNNs)** and **Recurrent Neural Networks (RNNs)**, play a key role in processing image and textual data, respectively.

- **CNNs** have become the standard for image-related tasks due to their ability to capture spatial hierarchies in data, making them ideal for visual recognition tasks [2]. These networks have been used extensively in text-to-image generation models to extract features from the image domain and guide the image synthesis process.
- **RNNs**, on the other hand, are used to model sequential data and have been employed in text-based tasks like caption generation and language modeling. When paired with attention mechanisms, they enhance the understanding of complex relationships between textual and visual features, leading to improved text-to-image generation [6].

2.2 Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) [1] have become one of the most important frameworks in deep learning for generative tasks. GANs consist of two neural networks: the **generator**, which generates synthetic data (in this case, images), and the **discriminator**, which evaluates whether the generated data is realistic by distinguishing it from real data. The two networks engage in a competitive process, improving each other iteratively until the generator produces high-quality, realistic outputs.

In text-to-image generation, **Conditional GANs (cGANs)** have proven particularly useful. cGANs allow the generation of images conditioned on specific input, such as a class label or textual description, making them ideal for tasks where both image and text data are required. For example, **StackGAN** [9] and **AttnGAN** [6] utilize cGAN architectures to generate high-resolution images from text descriptions.

2.3 The Role of GAN Variants in Text-to-Image Generation

Several GAN-based architectures have been proposed to improve the quality and diversity of generated images from text. One significant advancement is the **StackGAN** [9], which introduced a two-stage process: the first stage generates a low-resolution image, while the second stage refines the image to a high resolution. This approach enables the model to focus on global features first and then fine-tune the finer details.

- **AttnGAN** [6], another major contribution, incorporates an attention mechanism to focus on relevant parts of the input text while generating the corresponding image. This allows the model to capture fine-grained details in the generated images, making them more faithful to the textual description.
- **DM-GAN** [2] improves upon earlier models by introducing dynamic memory networks, which help refine the image generation process by retaining context and improving the semantic alignment between text and images.
- Additionally, newer approaches like **StyleGAN-T** [3] and **DF-GAN** [4] focus on scalability and improved image fidelity, allowing for controllable and high-quality text-to-image generation. These models incorporate techniques such as latent space manipulation and training stability improvements, further advancing the field.

2.4 Challenges in Text-to-Image Generation

 ©IJSRED: All Rights are Reserved

Despite the promising advancements, several challenges persist in the field of text-to-image generation. Key challenges include:

- **Semantic Gap**: Achieving high semantic alignment between text and generated images is difficult, especially when the text contains complex or abstract descriptions [2], [9].
- **Data Limitation**: High-quality paired datasets of images and text are required to train models effectively. The lack of large, annotated datasets remains a bottleneck in improving model performance [6].
- **Training Stability**: Training GANs is notoriously unstable due to the adversarial nature of the process. Models often face issues like mode collapse or vanishing gradients, which hinder the generation of diverse and high-quality images [1].
- **Evaluation Metrics**: Objective evaluation of generated images remains a challenge. Metrics like **Inception Score (IS)** and **Fréchet Inception Distance (FID)** [4], though widely used, may not always align with human perception of image quality, particularly when it comes to text-image alignment.

III. Text-to-Image Synthesis Overview

3.1 Problem Statement

Text-to-image synthesis is the task of generating a photorealistic image from a given textual description. This task is inherently complex due to the need for understanding and capturing the semantic meaning of the input text, along with the ability to generate detailed visual content. The challenge lies in bridging the gap between the high-level, often abstract nature of natural language and the concrete, spatially-organized content required for image generation [6].

Unlike traditional image generation tasks, which rely on structured input (e.g., class labels or image attributes), text-to-image synthesis must account for the full variability of natural language. Text descriptions can be vague, ambiguous, or highly detailed, and thus, models need to exhibit strong generalization and reasoning capabilities. Additionally, the generation process must retain consistency across the image, ensuring that the spatial relationships between different elements are coherent with the text description [9].

3.2 Applications

Text-to-image generation has a wide range of applications across multiple domains:

- **Art and Design**: Text-to-image models enable artists and designers to create visual content from textual prompts, thus streamlining the creative process and providing inspiration for new ideas [6].
- **Virtual Worlds and Gaming**: In video games and virtual environments, generating images or assets from textual descriptions can greatly enhance content creation, enabling more dynamic and diverse game worlds [3].
- **Medical Imaging**: Text-to-image models can assist in generating medical images based on textual descriptions, aiding in the training of diagnostic tools or even enhancing radiology and pathology reports [2].
- **Assistive Technologies**: For individuals with visual impairments, generating descriptive images from text could help create more accessible environments by visualizing content described in words [4].

These applications demonstrate the wide-ranging potential of text-to-image generation and underscore the significance of advancing this technology.

3.3 Challenges in Text-to-Image Synthesis

Despite its potential, text-to-image generation faces several significant challenges:

- **Ambiguity in Text Descriptions**: Natural language descriptions often contain vague or ambiguous information that can lead to multiple interpretations. This makes it difficult for models to generate images that align perfectly with human expectations [6], [9].
- **Fine-Grained Detail Generation**: Generating highly detailed images that

 ©IJSRED: All Rights are Reserved

match the input text at a fine-grained level (e.g., generating specific textures, object arrangements, and colors) remains a significant challenge. Models like **AttnGAN** [6] and **DM-GAN** [2] attempt to address this by focusing on attention mechanisms and memory networks, but further improvements are necessary.

- **Training Data Limitations**: The availability of high-quality paired datasets of images and text is essential for training accurate models. While datasets like MS-COCO and CUB-200-2011 have been widely used, they are limited in scope and often contain textual descriptions that are either too general or not sufficiently detailed for complex tasks [4].
- **Evaluation Metrics**: As previously discussed, evaluating the quality of generated images remains a challenge. Existing metrics like **Inception Score (IS)** and **Fréchet Inception Distance (FID)** [4] are effective for assessing image quality but fail to measure how well the generated images align with the textual description. This gap in evaluation frameworks necessitates more robust and nuanced metrics for assessing text-image coherence.

IV. Deep Learning and GAN-Based Models

4.1 Overview of GAN-Based Models

Generative Adversarial Networks (GANs) have been instrumental in revolutionizing text-to-image generation by offering a framework capable of generating high-quality images that align with textual descriptions. The core idea behind GANs is the adversarial setup, where two networks—the **generator** and the **discriminator**—compete against each other. The generator produces images from random noise (or conditional inputs), while the discriminator evaluates the authenticity of the generated images. Over time, this competition results in the generation of more realistic and visually coherent images [1].

To apply GANs to text-to-image generation, **Conditional GANs (cGANs)** are introduced, conditioning the generator and discriminator on textual information. This allows the model to generate images that are specifically tailored to the given text description. The challenge lies in ensuring that the generated image not only captures the overall content of the description but also reflects finer details in a coherent manner [6].

4.2 StackGAN: A Two-Stage Approach

One of the most influential models in the field of text-to-image synthesis is **StackGAN** [9], which introduced a two-stage generation process. In the first stage, a low-resolution image is generated from the text description, capturing basic shapes and structures. In the second stage, this low-resolution image is refined to a high-resolution version, adding finer details such as textures, colors, and finer spatial relationships.

This staged approach allows **StackGAN** to focus on global features initially and refine the image progressively, making it possible to generate high-quality images with a realistic appearance. The two-stage structure also mitigates some challenges in generating high-resolution images directly, which often results in blurry or unrealistic outputs.

4.3 AttnGAN: Leveraging Attention Mechanisms

Another groundbreaking model in text-to-image synthesis is **AttnGAN** [6], which incorporates an attention mechanism to focus on specific parts of the input text during the image generation process. This model enhances the quality of generated images by focusing on the most relevant regions of the text, allowing the generator to pay attention to fine-grained details and relationships between different objects in the description.

The attention mechanism in **AttnGAN** is particularly useful in scenarios where textual descriptions are complex or contain multiple entities. By attending to the most important parts of the description, the model generates images that are more coherent with the input text and exhibit better alignment with the intended semantic content.

4.4 DM-GAN: Dynamic Memory Networks for Improved Text-Image Alignment

DM-GAN [2] takes the concept of attention mechanisms further by introducing **dynamic**

 ©IJSRED: All Rights are Reserved

memory networks, which allow the model to better retain context and improve the alignment between text and generated images. The memory network helps store and retrieve important information during the image generation process, ensuring that the final output aligns more closely with the given textual description.

By utilizing dynamic memory, **DM-GAN** improves upon earlier GAN models that might overlook important details in the text, especially in longer or more complex descriptions. This approach enhances both the visual coherence and textual relevance of the generated images, making it one of the most effective GAN architectures for text-to-image synthesis.

4.5 Recent Advancements in GAN Models

Recent advancements have focused on improving the scalability and photorealism of generated images. Models like **StyleGAN-T** [3] aim to produce high-resolution images by incorporating techniques like latent space manipulation and improved training stability. These advances have led to the generation of images that not only have finer details but also exhibit high realism and flexibility in terms of generated content.

DF-GAN [4] is another recent model that builds on previous work by focusing on efficiency and scalability. By simplifying the GAN architecture and focusing on dynamic feature generation, **DF-GAN** can generate high-quality images faster and with fewer computational resources, making it a more practical option for large-scale applications.

4.6 Comparison of Models

In terms of architectural design, each GAN-based model offers unique strengths:

- **StackGAN** [9] excels at generating high-resolution images through a two-stage process, making it effective for capturing both global and fine-grained details.
- **AttnGAN** [6] uses attention mechanisms to improve the alignment between text and image, offering better control over the generated content.
- **DM-GAN** [2] introduces dynamic memory networks, improving the contextual understanding and semantic alignment between text and image.
- **StyleGAN-T** [3] and **DF-GAN** [4] have taken steps to enhance the scalability and quality of generated images, especially in terms of photorealism and computational efficiency.

Each of these models has contributed significantly to the field, addressing specific challenges such as image resolution, semantic alignment, and training stability.

V. Evaluation Metrics

5.1 Overview of Evaluation Challenges

Evaluating the quality of generated images in text-to-image synthesis is inherently challenging due to the subjective nature of visual perception and the complexity of aligning text with image content. Unlike traditional image generation tasks, where objective measures like pixel-level accuracy can be employed, text-to-image generation requires metrics that assess both the realism of generated images and their alignment with the provided textual descriptions [6]. This makes the development of effective evaluation metrics crucial for advancing the field and ensuring that models are not only generating realistic images but also faithfully representing the semantic meaning of the input text.

5.2 Commonly Used Evaluation Metrics

Several metrics have been proposed in the literature to assess text-to-image synthesis models. These can generally be divided into two categories: **image quality metrics** and **text-image alignment metrics**.

5.2.1 Image Quality Metrics

1. **Inception Score (IS)** [4]:

   The **Inception Score (IS)** evaluates the quality of generated images by using a pre-trained Inception network to classify images and measure their diversity. High IS scores are associated with images that are both realistic and diverse. While the metric has been widely used, it has been criticized for not measuring text-to-image alignment

 ©IJSRED: All Rights are Reserved

and for being influenced by the choice of pre-trained model.

2. **Fréchet Inception Distance (FID)** [4]:

   The **Fréchet Inception Distance (FID)** compares the distribution of features extracted from generated images to the distribution of features extracted from real images, using the Inception network. A lower FID indicates that the generated images are closer in distribution to the real images, which suggests high image quality and realism. This metric has become the standard for evaluating generative models, including text-to-image generation.

3. **Structural Similarity Index (SSIM)** [6]:

   The **SSIM** measures the structural similarity between two images. Unlike traditional pixel-level metrics, SSIM considers luminance, texture, and structure, providing a more perceptually relevant evaluation. However, it does not account for how well the generated image corresponds to the textual description.

5.2.2 Text-Image Alignment Metrics

1. **Recall@K and Precision@K** [9]:
   Recall and **Precision** metrics are often used to evaluate how well the generated images align with the textual description. **Recall@K** measures the percentage of relevant images retrieved in the top K generated results, while **Precision@K** measures the percentage of relevant images among the top K retrieved results. These metrics help assess the relevance of generated images to the input text.
2. **CIDEr** (Consensus-based Image Description Evaluation) [6]:
   CIDEr measures the degree of agreement between the generated image caption and reference captions. Though it is primarily used for image captioning tasks, it can be adapted to evaluate text-to-image generation by comparing the alignment between generated images and their textual descriptions. This metric emphasizes the semantic similarity between the description and the generated image, making it highly relevant for text-to-image synthesis tasks.
3. **BLEU (Bilingual Evaluation Understudy)** [2]:
   BLEU measures the n-gram overlap between the generated image captions and reference captions. While commonly used in machine translation and image captioning, BLEU has been used in the evaluation of text-to-image models to evaluate how closely the generated image descriptions match the provided text.

5.3 Subjective Evaluation

While quantitative metrics are important, they often fail to fully capture the human perception of image quality and text-image coherence. As such, **subjective evaluation** by human annotators remains an essential component in the assessment process. This involves evaluating the generated images based on:

- **Realism**: How photorealistic the image appears.
- **Text-Image Consistency**: How well the image matches the textual description.
- **Diversity**: The variety of images generated from different textual prompts.
- **Details and Fidelity**: The level of detail and accuracy in the image's content, such as textures, object shapes, and colors.

5.4 Challenges with Evaluation Metrics

Despite the availability of various metrics, several challenges remain in the evaluation of text-to-image generation models:

- **Text Complexity**: Text descriptions can vary in complexity, from simple phrases to detailed narratives, making it difficult for a single metric to account for all variations in text-to-image alignment [2].
- **Subjectivity in Human Evaluation**: Different evaluators may have varying standards of what constitutes "realistic" or "coherent" image generation. This subjectivity introduces variability in results.

 ©IJSRED: All Rights are Reserved

- **Discrepancy between Metrics and Human Perception**: As shown by studies comparing IS, FID, and human evaluations, automated metrics may not always correlate with human judgments of image quality or textual alignment, highlighting the need for improved metrics [6].

Conclusion

Text-to-image generation, driven by advancements in deep learning and Generative Adversarial Networks (GANs), has emerged as a transformative area in artificial intelligence, enabling the synthesis of realistic images directly from textual descriptions. Over the years, models such as StackGAN, AttnGAN, and DM-GAN have addressed critical challenges including text-image alignment, image quality, and diversity. Despite these advancements, significant hurdles remain—ranging from handling the ambiguity of natural language and achieving high-resolution outputs, to reducing computational complexity and improving evaluation methods. Future efforts must focus on refining attention mechanisms, enhancing memory networks, incorporating multimodal data, and adopting efficient model training strategies to push the boundaries of what these systems can achieve. As the field progresses, its applications are poised to impact creative industries, assistive technologies, and content generation platforms, underscoring the potential of AI to generate visually coherent, semantically accurate imagery from text with increasing realism and efficiency.

Refrences

[1] A. S. Rao, P. A. Bhandarkar, and P. A. Devanand, "Text to Photo-Realistic Image Synthesis using Generative Adversarial Networks," in *Proc. 2023 2nd Int. Conf. on Futuristic Technologies (INCOFT)*, Nov. 2023, pp. 1–6.

[2] M. Kang et al., "Scaling Up GANs for Text-to-Image Synthesis," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 1–10.

[3] A. Sauer et al., "StyleGAN-T: Unlocking the Power of GANs for Fast Large-Scale Text-to-Image Synthesis," *arXiv preprint arXiv:2301.09515*, Jan. 2023.

[4] Y. Liu et al., "A Framework for Image Synthesis Using Supervised Contrastive Learning," *arXiv preprint arXiv:2412.03957*, Dec. 2024.

[5] M. Kang et al., "Scaling Up GANs for Text-to-Image Synthesis," *arXiv preprint arXiv:2303.05511*, Mar. 2023.

[6] Z. Zhang et al., "ITI-GEN: Inclusive Text-to-Image Generation," in *Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, 2023, pp. 1–10.

[7] S. R. Mishra et al., "A Novel Photorealism Framework for Image Generation Using Generative Adversarial Networks," in *Proc. 2025 Int. Conf. on Artificial Intelligence and Machine Learning (AIML)*, Mar. 2025, pp. 1–6.

[8] J. Zhang, "A High-Resolution Image Synthesis Model Based on Conditional GANs," *Appl. Sci.*, vol. 15, no. 2, p. 706, Jan. 2025.

[9] M. Zhu et al., "DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-to-Image Synthesis," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1–10.

[10] H. Zhang et al., "StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks," in *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, 2017, pp. 5907–5915.

 ©IJSRED: All Rights are Reserved