SciNexuses

Journal Homepage: sciencesforce.com/scin

SciNexuses Vol. 1 (2024) 184–202

Paper Type: Review Article

Text to Image using Deep Learning: A Survey

Raghad Ahmed Gad [1], **Salma Saad Abdelshakour** [1] **and Ahmed Abdelhafeez** [2,3,*]

[1] Faculty of Information Systems and Computer Science, October 6th University 12585, Egypt;
Emails: Raghadahmed213@gmail.com; salmasaad.mohammad@gmail.com.

[2] Computer Science Department, Faculty of Information System and Computer Science, October 6 University, Giza, 12585, Egypt; aahafeez.scis@o6u.edu.eg.

[3] Applied Science Research Center, Applied Science Private University, Amman, Jordan; a_abdelhafeez@asrc.asu.edu.jo.

Received: 25 Aug 2024 **Revised:** 09 Dec 2024 **Accepted:** 29 Dec 2024 **Published:** 30 Dec 2024

Abstract

Text-to-image synthesis is an exciting marriage of natural language processing and computer vision for image synthesis from textual descriptions. This survey explores the discussed accomplishment of value in an industry that is rapidly evolving. Various attention mechanisms proposed by models such as AttnGAN have been discovered to improve fine-grained text-visual correspondences and hence deliver higher quality outputs. Comprehensive reviews of the text generation neural network have provided the base upon which various architectures and applications would be identified and investigated. Conditional GAN has defined how an image becomes a suitable image given a piece of text; methodological directions addressing reproducible human evaluation framework have established benchmarks for qualitative assessments of model performance. Semantic disentanglement methods also tackle the need for controlled generation, facilitating better interpretability and diversity. Bringing these developments together in one review, this survey discusses the issues now confronting researchers, such as computational complexity and consistency of evaluation, before describing the ways in which text-to-image generation will develop to enhance its academic and applied uses.

Keywords: Text-to-image Synthesis; Attention Mechanisms; Conditional GAN; Semantic Disentanglement; Computational Complexity.

1 | Introduction

The ability to create images from text description bridges the previously separate domains of natural language processing and computer vision; technology is a fundamental shift with deep-reaching consequences in a wide range of areas, from content generation to virtual reality and assistive technology. There has been unprecedented advancement in the field of generative modeling in Artificial Intelligence over the last few years; text-to-image synthesis has been a field of explosive expansion. Success in this field depends on advancements in neural networks, like training sophisticated models of neural architecture to establish mappings between the textual input meaning and the corresponding visual representations [6, 7].

Corresponding Author: aahafeez.scis@o6u.edu.eg

https://doi.org/10.61356/j.scin.2024.1518

Early text-to-image generation methods depended on either template matching or image retrieval techniques, where they would use pre-existing templates or scan databases to map textual queries to existing images [13, 14]. While effective for some use cases, these techniques were not creative and generative enough when it came to handling more intricate use cases. The introduction of generative models introduced a huge revolution in the field, spanning an extensive set of algorithms grounded on the principle of creative generation of images from text descriptions, including Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) [7, 15]. Conditional generative adversarial networks (cGANs), introduced in 2014, extended this concept by conditioning the image generation process on provided text inputs, thereby enhancing the resemblance between generated images and photogenic captions [9]. Much progress in the domain has made sense then. Stack GAN and AttnGAN have established the two-stage generation processes to refine the intermediate-generation low-resolution images into high-quality outputs to obtain higher resolution in the images [3, 10]. AttnGAN extends this concept further with the addition of an attention mechanism that allows the model to focus on specific words or phrases, potentially resulting in a better generation of visual representations that accurately reflect the finer details of textual descriptions [3]. These enhancements demonstrate a growing sophistication in the algorithms' ability to trade off visual realism and textual fidelity.

Together with these technological innovations, text-to-image model evaluation has likewise been a pertinent area of research. Quantitative metrics such as Inception Score (IS) and Fréchet Inception Distance (FID) are special metrics to quantify attributes of image and diversity [4, 11]. Nonetheless, these metrics are commonly criticized for their incapacity to explain subjective aspects such as contextual relevance and beauty [5, 11]. In this case, human evaluation frameworks have also been proposed to augment the automatic measures with a greater focus on reproducibility and establishing systematic criteria for evaluation [5].

This article reviews five seminal papers that have established the groundwork for text-to-image synthesis. AttnGAN enhances text-to-image alignment by employing attention mechanisms [3].

A systematic review of text generation models has been scrutinized; it is a wider view of the foundational techniques [2]. Conditional GANs are termed the backbone of text-to-image synthesis [4]. Reproducibility in model evaluation is approached with structured human assessment protocols [5]. Semantic disentanglement approaches are therefore presented, enhancing control over image attributes and generating more precise and interpretable outputs [1]. Despite considerable progress in addressing numerous problems, many problems continue to persist. The intricacy of high computation expenses, difficulty in precisely sensing fine text information, and subjective nature of evaluation are persistent limitations. Future research can focus on filling the gaps by studying hybrid models, complex and resilient models in natural language processing, and scalable evaluation models [6, 7].

Synthesized knowledge created here that indicates a general direction of trends aimed by this survey should provide both researchers and practitioners with valuable directions for the further development of text-to-image generation systems.

2 | Literature Review

It is required to introduce the most critical issues determining the development and challenges of text-to-image synthesis before introducing the literature review table. From the early approaches to the present ones, i.e., GANs, VAEs, and diffusion models, text-to-image models have made their development [7, 8]. In tasks such as computer vision, design, and content generation, where visual content is created by the machine from the text description provided, the models take a pivotal place [6, 9]. The difficulty still exists in processing large databases such as CUB and COCO to maintain image realism and text truthfulness and to keep in mind fine details [3, 5].

Some of the significant applications were found, such as StackGAN and AttnGAN, which improved image quality with multi-stage generation and attention, respectively [3, 10].

Scores like the Inception Score and human judgments, between the realism of generated images and similarity to input, are used in model assessment [5, 11].

The experiments in the Table 1 cover a wide range of methods, data sets, and results so that one can observe the merits and demerits of each method. This not only reflects present limitations but also indicates progress in the field [2, 6].

Table 1. Literature review.

Work/Year	Method	Dataset	Result	Strengths	Weakness
IEEE Transaction on Pattern Analysis and Machine Intelligence (TPAMI) in 2018	- (AttnGAN)‘ - (DAMSM)	CUB Dataset: Contains images of birds, COCO Dataset: It has more complex scenes	**Inception Score Improvement:** 1.14.14% Improvement*: On the CUB dataset. 2.170.25% Improvement*: On the COCO dataset	1. Fine-grained image Generation: The attention mechanism allows for detailed and high-quality image generation. 2. Improved Performance. 3. Multi-Stage Process	1. Complexity: The model's multi-stage process and attention mechanisms may increase computational complexity and training time. 2. Dependency on Datasets
IEEE Access in 2020	1. Vector-Sequence Models 2. Sequence-to-Sequence Models	COCO Dataset	1. Quality matrix 2. Performance matrix	1. Comprehensive Review: Review various deep learning approaches 2. Quality Metrics 3. Diverse Applications: Have multiple applications for text generation	1. Dataset Specificity: The paper does not delve deeply into specific datasets 2. Model Implementation*: It may not provide detailed implementation aspects
IEEE Transactions on Image Processing (TIP) in 2020	1. Semantics Disentangling Generative Adversarial Network (SD-GAN) 2. Siamese Network Structure 3. Semantic-Conditioned Batch Normalization (SCBN)	1.CUB-200 Dataset 2. MS-COCO Dataset	Achieved high performance on the CUB-200 and MS-COCO datasets for text-to-image generation.	1. Semantic Consistency 2. Detail Retention. 3. Innovation.	1. Complexity. 2. Generalization. 3. Evaluation Metrics.
International Conference on Machine Learning (ICML 2016)	1. Generative Adversarial Networks (GANs) 2.Text-to-Image Generative Network	1.Oxford-102 Flowers:8189 image 2.CUB-200 Birds: have 11788 images. 3. COCO Dataset: eighty thousand images	The method successfully synthesized images from textual descriptions, with better quality and diversity than previous approaches.	1. Innovative Use of GANs. 2. Multimodal Learning. 3. Improved Image Quality.	1. Image Resolution. 2. Diversity of Generated Images. 3. Dependency on Text Quality
IEEE (2021)	1. Human Evaluation Protocol 2. statistical Analysis 3. Crowdsourcing	1. Varied Datasets 2. COCO 3. CUB	Improved Reproducibility: By establishing clear protocols. Benchmarking: Method tested and validated on models	1. Reproducibility. 2. Rigorous Evaluation. 3. Applicability.	1. Complexity of Human Evaluation. 2. Dependence on Crowdsourcing. 3. Resource Intensive.

3 | Evolution of Algorithms in Text-to-Image Generation: From Early Methods to Modern Deep Learning Approaches

3.1 | Early Approaches: Image Retrieval and Template Matching

Before the development of deep learning, text-to-image systems created images from text descriptions using image retrieval techniques, rule-based systems, and template matching.

- Template Matching: It employs existing picture templates that are associated with terms or concepts. The simple sentence "a red apple" would use the term red as per the defined meaning and create an existing image of an apple. The technique was, however, constrained by the amount that existed in existing templates and was not highly flexible.
- Image Retrieval: Rather than creating images, the initial image retrieval systems were trying to find existing images in the database that were visually equivalent to the text query. Rather than creating new images, the systems wanted to obtain existing ones. An indexed database would be queried with the text words to get visually equivalent entries.

3.2 | Generative Models: GANs (Generative Adversarial Networks)

When Generative Adversarial Networks, or GANs, were first proposed by Ian J. Goodfellow and others in 2014, they were an overnight sensation in the machine learning world. GANs are an unsupervised machine learning task that is a duo of neural networks by the name of a generator and a discriminator that compete to observe, record, and replicate shifts in a dataset [14].

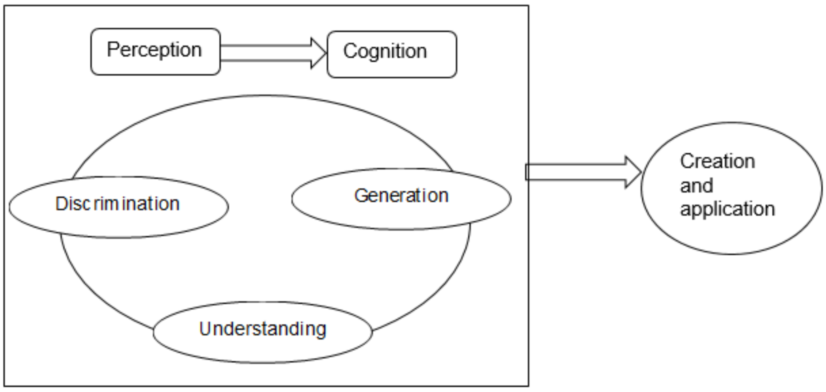

Figure 1. The main phases of GAN [6].

- **Generator**:

The generator in GANs is a neural network that creates false data to train the discriminator as shown in Figures 1-3. The generator takes as input the fixed-size random noise vector and produces a sample. The synthetic samples are adverse training samples of the discriminator.

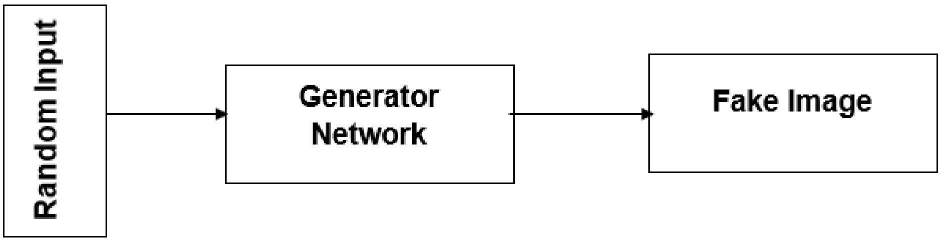

Figure 2. Generator [6].

In a generative adversarial network (GAN), there is a single objective of the generator, and that is to generate outputs that are recognized by the discriminator as real. The generator is trained on a noisy input vector and transforms it into a data instance. It involves the following three processes:

A generator network generates artificial data by the acceptance of random input. A discriminator network assigns real or synthetic labels to the data generated. If, in this case, the generator fails to fool the discriminator, a generator loss function penalizes the generator. From observing how each weight is acting towards contributing to the output, backpropagation fine-tunes the weights in the network. Additionally, backpropagation calculates the gradients possible, used to update generator weights so they are best optimizing it in the long term.

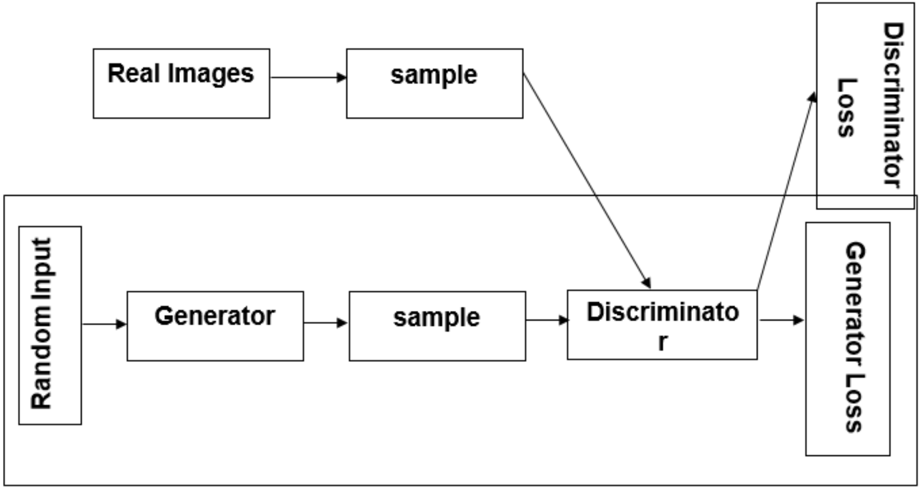

Figure 3. Generator Training [7].

- **Discriminator**

The discriminator is a neural network structure that is tasked with distinguishing between real and artificial data produced by the generator as shown in Figure 4-5. The training data has two origins: real data samples, such as images of human faces, birds, and banknotes, used as positive samples, and artificial data samples generated by the generator and used as negative samples.

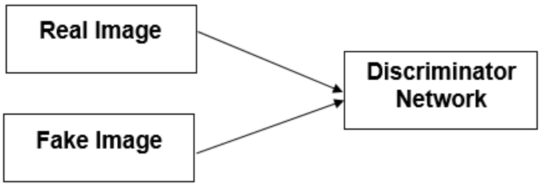

Figure 4. Discriminator [8].

The discriminator usually employs two loss functions in training but does not pay attention to the generator loss and focuses only on the discriminator loss.

Identifying genuine and artificial data produced by the generator is the work of the discriminator during training.

When the model labels actual samples as imitations or vice versa, discriminator loss is used to penalize the model. The discriminator uses the propagated loss across its network to do backpropagation and update its weights so that it learns.

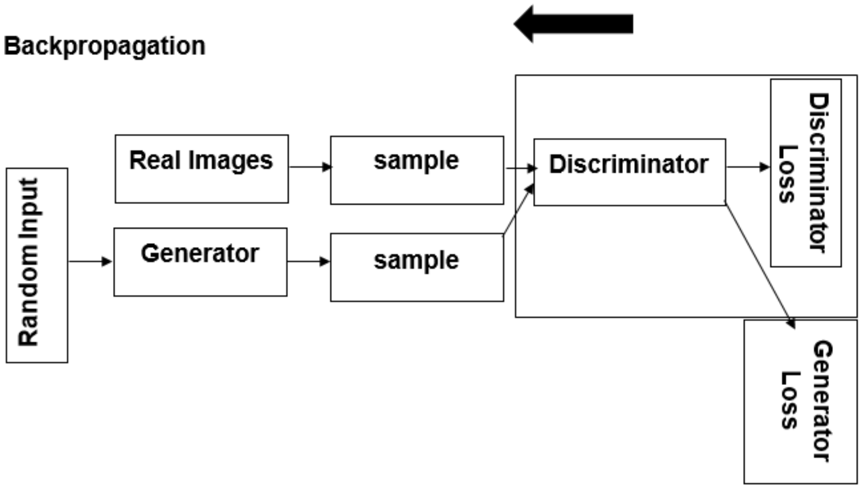

Figure 5. Discriminator Training [9].

-The mathematical equation for Gan can be represented as:

$$V(D,G)=Ex\sim Pdata(x)[logD(x)]+Ez\sim p(z)[log(1-D(G(z)))] \quad (1)$$

Where:

- $G = Generator$
- $D = Discriminator$
- $p(z) = distribution\ of\ generator$
- $Pdata(x) = distribution\ of\ real\ data$
- $x = sample\ from\ Pdata(x)$
- $z = sample\ from\ p(z)$
- $D(x) = Discriminator\ network$
- $G(z) = Generator\ network$

- **Conditional GANs**

One of the GAN model variants, the Conditional Generative Adversarial Network (cGAN), was proposed by Ian Goodfellow in 2014. It provides the output with the ability to be conditioned on a text or a label such that the generation would be controllable. It has been named a conditional GAN because of this capability as shown in Figure 6.

- Challenges: Although there was an improvement, cGANs were unable to generate intricate images and generated unrealistic or blurred images instead. They also failed to embed the text description details.

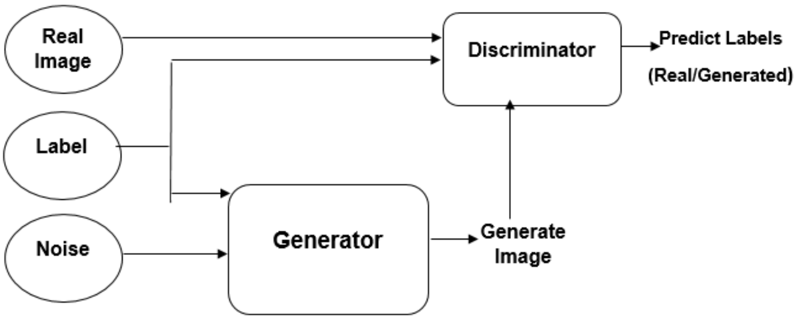

Figure 6. Conditional GANs [7].

Conditional GANs

By adding other information, denoted as y, to the discriminator and the generator, one can generalize the model to a conditional model. Class labels and other modality information are only two among the myriad types of auxiliary data that can be included above. By concatenating y with the final input noise of the generator, z, a standard hidden representation is created.

Generator Architecture

The generator takes as input the previous input noise, represented by z, and usually additional data, represented by y. They are combined into one shared joint latent representation, from which the conditional synthetic samples for the input are sampled. This implicit representation allows for flexibility during adversarial training.

Discriminator Architecture:

Both the original data (x) and the extra information (y) are provided to the discriminator. It must be able to distinguish between the actual data and the data generated synthetically by the generator.

Loss Function:

$$GminDmaxV\ (D, G) = Ex{\sim}Pdata(x)[logD(x|y)] + Ez{\sim}p(z)[log(1-D(G(z|y)))] \qquad (2)$$

Where:

- E: *Represents the expectation operator, which is expected value of a random variable*
- Ex ~ Pdata(x) Represents the expected value concerning the real data distribution Pdata
- Ez ~ p(z): Represents the expected value concerning the prior noise distribution

3.3 | StackGAN (Stacked Generative Adversarial Networks)

A StackGAN stacks two GANs on top of each other sequentially to generate high-resolution images as shown in Figure 7. Compared to vanilla GAN, which attempts to construct a picture in one manner but never high-definition or quality, this is an improvement. The two-step process StackGAN provides is that there is a second conditional variable that introduces information the generation process uses to guide it. Following a text embedding, Stage I of the initial GAN generates a low-resolution image and learns about shapes and colors. The second GAN in Stage II then refines the image with fault correction and the inclusion of finer details and generates a high-resolution, photorealistic image. With the help of these two stages together, StackGANs work much better than vanilla GANs and conditional GANs [3, 7].

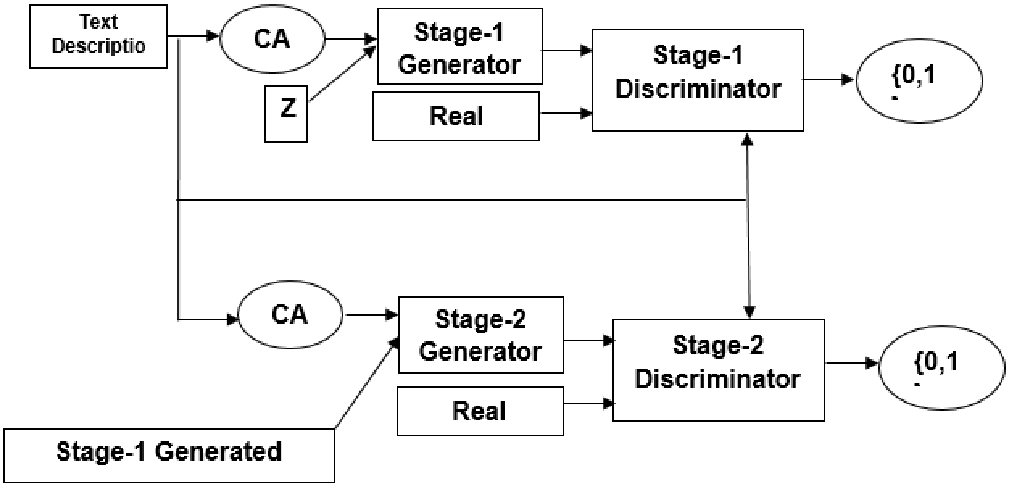

Figure 7. StackGAN [3, 7].

3.4 | AttnGAN (Attention Generative Adversarial Network)

AttnGAN was presented in 2018. AttnGAN or Attention Generative Adversarial Network is an improved model as compared to other models like StackGAN and traditional GANs. AttnGAN particularly handles generating high-res images from textual inputs with the help of attention. The Attentional Generative Adversarial Network begins from a low-res rough image at the start and enhances it one step at a time to generate a final image. Then the Multi-modal loss of attention helps the model to focus on certain words or parts of the text to generate different regions of the image. It produces a significantly better alignment between the generated image and the content in the text description. Use an attention-based approach where, apart from using the whole text description as equivalent [3, 4], it learns to concentrate on the most relevant words when generating image regions.

3.5 | Evaluation of GAN

The evolution of GANs does not end at AttnGAN (2018). DM-GAN (Dynamic Memory GAN) came out in 2019, with improvements on AttnGAN, through the introduction of a dynamic memory module. The module allowed the model to remember and refine complex details of complicated images during image generation, thereby improving the final image. Released in 2019 as well was Control GAN, where a self-attention mechanism introduced further assisted in controlling image generation. This capability enabled Control GAN to achieve a stronger correspondence between certain textual descriptions and the respective regions in images, resulting in significantly more control over the generated content using this approach [13].

4 | Variational Autoencoders (VAEs)

VAEs are a second type of generative approach that was utilized for text-to-image translation, but not as much as was true for GANs.

- Process: The input text is encoded into a latent space in VAEs, and an image is reconstructed using the latent code. The VAE model learns a conditional image distribution given the input text.
- Strengths and Weaknesses: VAEs have a more organized latent space than GANs and are therefore suitable for image manipulation. They generate less detailed and nevertheless sharper images than using GAN-based techniques.

5 | Diffusion Models

Diffusion models form a class of generative models that build high-fidelity data, such as images, by a diffusion-driven physical process as shown in Figure 8. Diffusion models work by successively adding noise to an image and learning how to invert the process to synthesize new clean images from noise.

- Noise Addition: also called the forward process it starts with a clean image and is progressively adding noise in several steps until it is purely noise. This part is a training step, where the model is trained to debase images.
- Noise Removal: also known as the inverse process, the model is trained in the inverse process of adding noise. Starting with random noise, step by step, it removes noise and generates a final image according to the data it has been trained on or the text under which it has been guided [7, 15].

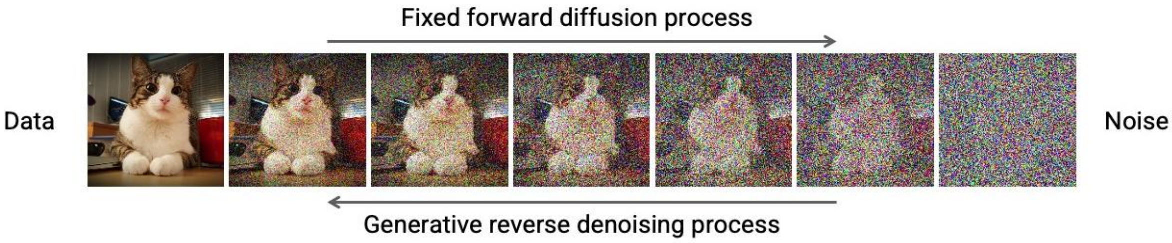
Fixed forward diffusion process
Data
Noise
Generative reverse denoising process

Figure 8. Diffusion models [7, 15].

6 |Transformers and Large-Scale Language Models

It went through development into more sophisticated versions, for example, DALL·E 2 and DALL·E 3. Later developments brought dramatic enhancement in image quality as well as in the comprehension of complex textual prompts [16].

6.1 |DALL•E 2

in 2022, replaced this with a significant architecture change from its previous version, changing from the VAE architecture to a diffusion model. Diffusion models produce images by progressively purifying a random noise input into a coherent image. Therefore, this led to a visible enhancement of output sharpness and detail compared to VAE-based models. Thus, the switch enabled DALL·E 2 to create images of better resolution that are more photorealistic without sacrificing a high degree of fidelity to the text description. In addition, the model was fine-tuned concerning the text via the addition of CLIP guidance on image generation, thereby improving its ability to accurately depict complex or abstract environments based on the provided description [7, 15].

6.2 |DALLE 3

released in 2024, further enhanced the functionalities of DALL·E 2 to also include natural language interaction like ChatGPT. Users could give longer, multi-step instructions for the image generation process with greater levels of artistic freedom and high-resolution specificity. DALL·E 3 was even more capable of creating more realistic and subtle images, with more subtlety around the detailed representation of objects, scene composition, and general artistic style [17].

7 |Imagen (Google)

Google in the year 2022 produced Imagen, a state-of-the-art text-to-image model, and the newest benchmark for image generation quality. As with DALL•E 2, Imagen employs one of the most powerful techniques available for generating images from random noise to high-resolution, detailed images: diffusion models. Diffusion models typically begin with a pattern of random noise and advance through iterative steps, systematically "denoising" it while leveraging the input text prompt to direct the process toward the target result in image generation.

The difference in Imagen is that it focuses on the intrinsic understanding of language. Unlike earlier models, including DALL•E and AttnGAN, which made heavy use of the process of mapping text tokens to visual features, Imagen used Google's pre-trained T5 language model to develop an in-depth understanding of the semantics involved in rich and subtle textual instructions. This helped the model to interpret more complex, abstract, or refined instructions than their earlier counterparts had done [15].

How it works

Imagen first passes the input text prompt through the T5 language model. This allows the model to understand the complete implications of a description, along with the subtleties of word order contextual hints, and implied meaning. Due to this, it can form a very deep understanding of complicated and abstract prompts.

It is performed with a diffusion model following text input encoding. It begins with the dirty image initialized randomly and progresses through a sequence of denoising rounds stepwise progressively towards smooth noise into the shape of perceiving some meaning. With every iteration of this denoising, the model is trained on the T5 encoded text representation and hence ensures that the generated image always remains in agreement with the input prompt's progress.

Imagen also uses multi-scale image generation techniques. It initially creates a low-resolution image that grasps the overall structure and general composition of the provided prompt. The model then gradually refines the low-resolution image through iterations, with each phase adding detail, texture, and overall correctness, producing a high-resolution, highly detailed image adhering to the text representation [15, 16].

8 | Stable Diffusion Model

Stable Diffusion is an open-source text-to-image generative model, released by Stability AI in 2022, that creates high-resolution images from a text prompt out of a diffusion model. Some of the best models include DALL•E 2 and Imagen. Stable Diffusion is unique in that it brings the diffusion model with an unpretentious and accessible version that can be run efficiently on typical consumer-level hardware. This is a field where developers and researchers can create stunning images with minimal utilization of computing resources [15].

How It Works:

Stable Diffusion is based on what is called a Latent Diffusion Model, which generates in latent space instead of in pixel space directly as shown in Figure 9. LDMs do not need to deal directly with high-dimensional, computationally expensive pixel data, but instead represent images using a pre-trained encoder, a neural network such as a Variational Autoencoder, or another similar type of architecture. This therefore enables the model to operate better on a particular instance and preserve valuable characteristics of the image [7].

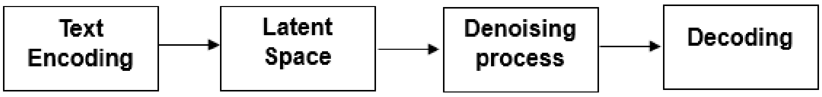

Figure 9. How stable diffusion works.[7]

The overall architecture of Stable Diffusion starts with a text encoder that can take input in the form of a text prompt. The input is processed by using a pre-trained language model, usually CLIP, to embed the semantic meaning of the text. By operating in latent space, a compressed space representation of the image level of efficiency is achieved through a reduction of the dimension of the space into which random noise is added. The operation starts in the latent space with an addition of random noise, which is gradually diminished at each subsequent step in the guidance of a specified encoded text prompt, thereby continually refining the image to better reflect the provided description. Once this noise is minimized, the latent representation is converted back into pixel space to generate a super-resolution image closer to the input text description [7].

9 | Parti Model

There is another advanced model from Google Research called Parti - for Pathways Autoregressive Text-to-Image new, unnatural/text generation genre as shown in Figure 10. It differs from diffusion models such as DALL•E 2 and Stable Diffusion since it relies on an autoregressive transformer architecture, which is characteristic of natural language processing activities such as text generation. The key distinction is that Parti does this autoregressive process on image generation, [15] breaking down images into discrete token sequences-partly as words are composed into sentences and then generates the image token by token based on the input text prompt [17].

How It Works:

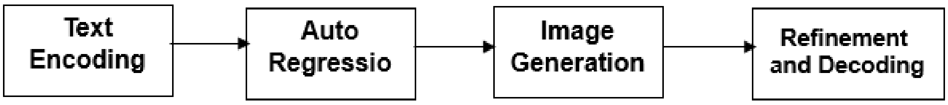

Figure 10. How Parti works [15].

First, Parti encodes a given text prompt using a transformer-based language model that captures the semantic meaning of the text guiding the image generation process. Then, it converts the image into a sequence of discrete tokens using a discrete VAE or any similar tokenizer. Parti predicts this sequence of tokens autoregressively, where the image is generated one token at a time conditioned on all previously generated tokens, like models like GPT, which output a word based on the context of words coming before it. When this sequence of tokens that represent the image is complete, it decodes back into pixel space to form the final image. This autoregressive process ensures a high degree of coherence between the image and the text prompt of the details involved in highly specific and accurate visual outputs.

10 | Muse Model

MUSE is an advanced model from Google Research for text-to-image generation introduced at the end of 2022 as shown in Figure 11. It works via a masked generative transformer approach that enables image generation at extremely high efficiency and with detail. While diffusion models synthesize images iteratively from noise, MUSE follows an autoregressive approach: It predicts masked areas of an image in parallel, like how language models would predict missing words. It focuses on diversification and faithfulness in images, generating full-of-detail and contextually accurate images from text prompts. It is versatile in creative and editing tasks related to inpainting, outpainting, and even mask-free editing. Masked modeling also makes Muse efficient for creating high-resolution images in fewer steps relative to diffusion-based models, hence offering faster performance with high-quality outputs [16].

How It Works:

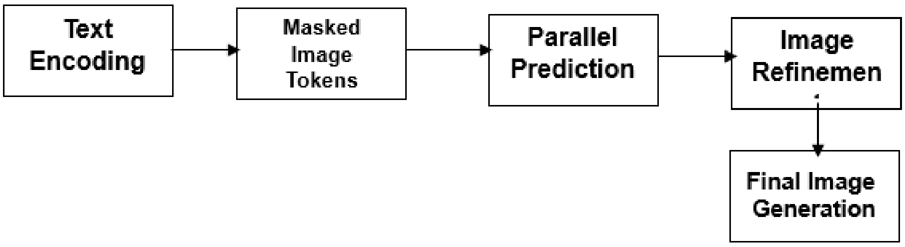

Figure 11. How Muse works [16].

This goal is achieved by making a preliminary analysis of the input text prompt to find meanings that provide a basis for the user's wishes for the image generated. Rather than interacting directly with pixel-based images, Muse splits such images into sub-units called tokens, which are like jigsaw puzzle pieces. The model masks or conceals some parts of the image and then predicts the missing tokens based on the text prompt. In contrast to step-by-step image generation diffusion models, Muse predicts all missing tokens simultaneously, so it can process different regions of the image in parallel. Through numerous iterations, the model refines the image to achieve greater text prompt consistency when filling in the tokens. Once it completes predicting and refining all the tokens, Muse reconstructs the image to generate a high-quality output that is aligned with the user's description.

11 | Recent Trends and Future Directions

- Control and Interactivity: Models are being developed in a way that gives users greater control over the generation process. This includes editing certain parts of the generated image (e.g., style, color) and developing more interactive interfaces.

- Multimodal Models: As multimodal models are becoming more necessary, there need to be models that not only generate images from text but also operate in other modes, i.e., generating video or 3D from text.
- Efficiency and Accessibility: Future work is aimed at making the training and deployment of such models cheaper computationally without sacrificing image quality. Quantization and knowledge distillation are some of the methods being tried out for making models like DALL-E 2 and Stable Diffusion more accessible to the masses.

Before the algorithms. Table 2, briefly state the history of text-to-image generation models:

- Before 2010, templates or keyword-based search but non-generative methodologies were used for template matching and image searching that were not as creative.
- Generative Models: VAE and Conditional GANs in 2013–2014 provided the ability to utilize generative abilities but not so acuity, resolution, and subtlety in the image. Advanced GAN Models: StackGAN (2016) and AttnGAN (2018) improved image quality through multi-stage generation and attention mechanisms with more complexity and computation.
- Recent Advances: DALL-E and Diffusion Models (2021) brought transformers and denoising processes to the forefront, which allowed for high-quality, imaginative results but at the expense of tremendous resource usage.

New models like Imagen, Stable Diffusion, and Deep Floyd IF (2022–2023) offer higher fidelity, higher customizability, and higher efficiency, but fiddly prompts and high compute costs are still problems.

Table 2. Text-to-image generation algorithms [7-8, 15].

Algorithm /Model/Year	Key Technique	Description	Strengths	Limitations
Template Matching Pre-2010	Rule-based, pre-designed templates	Uses predefined image templates that correspond to specific text descriptions	Simple, easy to implement	Lacks flexibility, limited by template availability
Image Retrieval Pre-2010	Retrieval-based	Retrieves existing images from a database based on text keywords	Finds high-quality real images	Not generative, limited creativity
Variational Autoencoders (VAE) 2013	Autoencoder-based latent space generation	Encodes text into a latent space and decodes it into an image	Structured latent space, easy interpolation	Less sharp images compared to GANs
Conditional GAN (GCN) 2014	GANs conditioned on text	Uses GANs to generate images conditioned on text descriptions	Can generate new images based on input text	Struggles with minute details, blurry images
StackGAN 2016	Two-stage GAN	Generates low-resolution images first, then refines them to high-resolution	Improved image resolution and detail	Computationally expensive, still some artifacts
AttnGAN 2018	Attention mechanism with GAN	Uses attention to focus on specific words while generating different image regions	Captures fine-grained details from text	Complex training process
DALL-E 2021	Transformer-based architecture	Uses transformers to model relationships between text and image pixels	Generates creative, novel combinations	Requires large datasets and computer power

Diffusion Models **2021**	Denoising process	Gradually denoises a noisy image into a clear one, conditioned on text	High-quality, photorealistic images	Computationally expensive, slow generation
Imagen **2022**	- Diffusion Models: Refine noise into images. - Large Language Models: Enhance text understanding. - Conditional Generation: Generate images based on text prompts.	A text-to-image model by Google Research that creates high-quality, photorealistic images from textual descriptions using diffusion processes.	- Produces high-fidelity, photorealistic images. - Strong comprehension of complex prompts. - Generates diverse outputs from the same text.	- Requires significant computational resources. - Image quality varies with prompt wording. - Dependent on the quality of training data.
Stable Diffusion **2022**	Latent Diffusion, VAE (Variational Autoencoders)	Stable Diffusion is a highly popular open-source model for text-to-image generation.	Open-source, highly customizable, low resource consumption, and large community support.	Struggles with detailed or complex prompts sometimes create artifacts in images.
Parti **2022**	Autoregressive model, Staged Diffusion	Parti (Pathways Autoregressive Text-to-Image) by Google generates images via multiple stages.	High image fidelity and diverse image generation.	Slow due to multiple stages; high computational cost.
Muse **2023**	Transformer-based architecture, Autoregressive Model	Muse is a text-to-image diffusion model from Google focused on generating high-quality images.	Fast generation, highly efficient, and lower computational cost.	Still experimental and lacks a diversity of trained subjects.
DeepFloyd IF **2023**	Multi-stage diffusion, Text conditioning via large model	Deep Floyd IF is an advanced multi-stage diffusion model designed for detailed text-to-image output.	An elevated level of image fidelity, and ability to generate intricate and creative visuals and detailed images.	Requires extensive computational resources and limited availability for public use in comparison to others.

12 | Methodologies Comparative Analysis

12.1 | AttnGAN: Fine-Grained Text to Image Generation

AttnGAN multi-stage that utilizes attention mechanisms to generate high-quality images. The primary aim of AttnGAN's method is to emphasize important parts of the text description during image generation. With this emphasis, the model can generate images that are coherent and detailed in their information.

Strengths:

- Generating contextually appropriate and detailed outputs.
- The multistage method enhances the quality of the images.

Weaknesses:

- Highly computationally intensive, demanding extremely high resources for training and inference.

12.2 | A Systematic Literature Review on Text Generation

This work does not suggest a novel algorithm; rather, it offers an exhaustive survey of current methodologies within the field of text generation. The article classifies deep learning models, i.e., recurrent neural networks (RNNs), Long Short-Term Memory (LSTM) networks, and Transformers, with their relative strengths and weaknesses in producing coherent textual output.

Strengths:

- Gives a general overview of many models.
- Sets the significance of metric measurement.

Weaknesses:

- Lacks an innovative algorithm.

12.3 | Generative Adversarial Text to Image Synthesis

This paper introduces a Conditional GAN (cGAN) Algorithm for text-to-image synthesis. The generator and discriminator of this algorithm are conditioned on text so that synthesized images are strongly associated with input text descriptions.

Strengths:

- The conditional GAN architecture handled the models' intricate text-image relationships well and produced high-quality results.

Weaknesses:

- Must be fine-tuned very well for the best performance.

12.4 | Toward Verifiable and Reproducible Human Evaluation

This study changes direction toward the assessment of text-to-image generation models, as opposed to their generation techniques. The authors propose a Human Evaluation Framework directed at systematic image quality judgments along the dimensions of clarity, relevance, and overall coherence concerning the textual input.

Strengths:

- Establishes the basis for human evaluation in favor of clarity, relevance, and reproducibility.

Weaknesses:

- Favored testing as opposed to production.

12.5 | Semantics Disentangling for Text-to-Image Generation

This paper introduces an approach based on Disentangled Representation Learning, a conditional image generation under the control of some attributes such as color, shape, and texture. The authors modify the generic GAN framework to accommodate such disentangled representations.

Strengths:

- This paper proposes the idea of disentangled representations to enable explicit control over features in synthesized images.

Weaknesses:

- The complexity of disentangled representation learning can be difficult to train and incorporate into existing architectures.

State these points briefly before providing the table:

The Table 3 shows some of the most important contributions in text-to-image synthesis: attention mechanisms, semantic disentangling, etc.

Other algorithms highlight methods that are beneficial to enhance the image quality as well as the text alignment.

While these techniques have advantages, such as improved interpretability and diversity, they are afflicted by problems such as computational complexity and mode collapse.

Human evaluation models offer complex analysis and reproducibility of output models. Limitations exist despite improvement, for example, computationally prohibitive cost and infuriating tuning.

Table 3. Comparison of key algorithms and methodologies used in each of the five papers [3-5].

Paper Title	Algorithm/Method	Key Features	Strengths	Limitations
AttnGAN	Attention-based GAN	-Multi-stage generation - Text encoding By RNN -Fine-grained attention	- High-quality image synthesis - Better alignment with text	-Computationally intensive - Complexity in tuning
Systematic Literature Review on Text Generation	Overview of Deep Learning Models	- Analyze RNN, LSTM, Transformer, GAN - Various evaluation metrics	- Comprehensive understanding of models and applications	- Does not propose a new model
Generative Adversarial Text to Image Synthesis	Conditional GAN (GCN)	- Text embedding as conditioning input - Dual training of generator and discriminator	- Captures complex text-image relationships - High image diversity	- Can suffer from mode collapse
Toward Verifiable and Reproducible Human Evaluation	Evaluation Framework	- Structured human evaluation - Guidelines for reproducibility	- Rigorous assessment of model outputs - Clear evaluation protocols	- Focuses on evaluation, not generation
Semantics Disentangling for Text-to-Image Generation	Disentangled Representation Learning	- Semantic component separation - Enhanced GAN architecture	- Improved interpretability - Controlled image attributes	- Complexity in representation learning

13 | Challenges and Limitations

The landscape of text-to-image generation is changing very quickly, and there are nonetheless a few central challenges that persist:

- Quality of Generated Images: Although newer models have worked towards the realism of the generated images, there still exist some situations where generated images may be of inferior quality or fidelity. These issues range from blurriness, artifacts, and insufficiency of detail, and can contribute to usability loss of generated images.

- Evaluation Metrics: The issue of measuring generated images still eludes us. Existing measures such as Inception Score (IS) and Fréchet Inception Distance (FID) have their shortcomings and do not accurately capture the qualitative aspect of images. Human judgment, while useful, is unreliable and variable, and one cannot infer model performance.

- Understanding Context: The current models may not be able to understand context or subtleties in written definitions. This can lead to image misinterpretation, with the generated image failing to carry the intended meaning as planned by the text.
- Scalability and Efficiency: Most advanced models need enormous computational resources to train and make inferences, and this may inhibit their availability and usability in real-time systems.

To counter these challenges, some of the potential directions to explore and create are:

- Enhanced Model Architectures: Future work can be directed toward creating new architectures with hybrid strategies by combining the strategies of different models (e.g., transformers, GANs, and attention mechanisms) to enhance the quality and diversity of generated images.
- Advanced Evaluation Techniques: Enhanced evaluation techniques, both qualitative and quantitative, are required. More recent metrics with the ability to measure improved subjective quality of output images and coherence of text description would be beneficial.
- Incorporating Contextual Understanding: Experiments can examine in what way the contextual knowledge of models and the text richness can be increased. It can be achieved by utilizing more diverse quantities of data or by having better natural language processing such that the text is better represented.
- Data Efficiency: Explore ways in which models can be efficiently trained from scarce data, i.e., few-shot or zero-shot learning strategies, for the generalization and popularization of text-to-image technology.
- User-Controlled Generation: Subsequent work may include allowing users greater control of generation. Models can generate closer-to-user taste and specification through the introduction of user-specified parameters or feedback loops.
- Cross-Modal Learning: Inspiration from cross-domain methods, such as vision-language pre-training, can be used to enhance coupling between the text and image modalities to support better understanding and generation.

14 | Evaluation Metrics in Text-to-Image Generation

i) Inception Score (IS):

Originally, among the evaluation metrics that were designed to estimate the quality of the produced image. It has two metrics: confidence and diversity [17].

- Confidence (sharpness of prediction):

  Precision of pre-trained model to label every generated image in a specific class.

- Diversity (range of generated images):

  Diversity is a term used to say how different the images that are being generated in set in total, if the model passes images to a great number of classes, then diversity will be high and conversely if the model passes generated images to one class, then diversity will be low. An increase in diversity means that the model is good.

ii) Fréchet Inception Distance (FID):

It is a more advanced metric that compares real and generated image statistics by computing the Fréchet distance between the feature representations of the two sets. Sees image quality and diversity. It is superior to the inception score as it directly compares the generated image with the real one.

iii) Recall/Precision:

- Precision: measures how realistic and pretty the output images are. It considers how good-looking the images are but not how close they are to the input text.
- Recall is calculated by measuring how close the produced image is to the text.

iv) Human Evaluation:

The process through which human beings need to analyze the quality of images generated by models based on provided criteria. Since machines cannot always comprehend details.

v) Runtime and Computational Resources:

Specify the computation time and the computation utilized to generate and test images in text-to-image models. Since models will need time and memory to provide accurate outputs.

vi) Challenges in Evaluation:

Evaluation of text-to-image generation models has some challenges:

- Limited Metrics: No single metric can represent all aspects of image relevance and quality. Quantitative and qualitative metrics are both necessary for proper balance.
- Context Sensitivity: Images generated could be context-dependent, i.e., a model could be good in one context and terrible in another. Evaluations must consider numerous contexts to give robustness.

Give the following key points before showing the Table 4:

Image generation performance must be measured; quantitative metrics at various levels that determine the quality and relevance of image generation models from text are given.

- Objective-Subjective Evaluation: IS and FID, computational, and Human Evaluation, qualitative.
- Trade-off: Some measures will give speed, but these are superficial, and others that are more significant are computationally costly.
- Challenges: All the proposed measures to date are missing something; some are biased, some are computationally intensive, and some need a lot of data.
- Human Evaluation: Even though it is the most helpful to date, this is extremely time-consuming and difficult to scale.

Table 4. Comparison of the evaluation metrics [5, 11].

Metric	Description	Strengths	Weaknesses	Limitations
Inception Score (IS)	Measures the quality and diversity of generated images using a pre-trained Inception model.	- Simple to compute - Useful for quick assessments	Does not compare to real images and is sensitive to class labels.	- May favor certain classes - Not fully indicative of image quality
Fréchet Inception Distance (FID)	Measures the distance between feature distributions of generated and real images.	- Robust and dependable - Accounts for image quality and diversity	Slower to compute and needs a large real dataset.	- Sensitive to the choice of features - Requires many samples
Runtime and Computational Resources	Assesses time and resources for image generation.	Ensures efficiency and scalability.	Does not measure quality or diversity directly.	Focus on efficiency, not long-term performance.
Recall/Precision	Evaluate image quality (Precision) and diversity (Recall).	Distinguishes between quality and diversity	Complex to compute and needs large datasets.	High computational cost, data dependent.

Human Evaluation	Involves human judges assessing relevance, clarity, and overall quality of images	- Captures subjective quality - Provides nuanced insight	Time-consuming, subject to bias.	- Subjective and variable - Time-consuming and labor-intensive - Hard to scale;

15 | Conclusion

In total, this survey has outlined five primary contributions to the development of a text-to-image generation model. Every paragraph, from AttnGAN attention modules to why human evaluation is necessary, and which constitutive pieces can be unbundled, provides the precise piece of the puzzle to advance this field forward. Although breakthrough jumps are happening for realistic and diverse image generation from text-based data, model complexity, evaluation, and scalability problems do need additional effort. But these can be overcome and improved, and robust text-to-image generation systems can be proposed such as hybrid models and multimodal fusion. This paper proposes to address some of the research needs both in terms of modern technology and improvement of existing text-to-image approaches and image generation systems.

Acknowledgments

The author is grateful to the editorial and reviewers, as well as the correspondent author, who offered assistance in the form of advice, assessment, and checking during the study period.

Author Contribution

All authors contributed equally to this work.

Funding

This research has no funding source.

Data Availability

The datasets generated during and/or analyzed during the current study are not publicly available due to the privacy-preserving nature of the data but are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that there is no conflict of interest in the research.

Ethical Approval

This article does not contain any studies with human participants or animals performed by any of the authors.

References

[1] G. Yin, B. Liu, L. Sheng, N. Yu, X. Wang, and J. Shao, “Semantics Disentangling for Text-to-Image Generation,” vol. 1.

[2] N. Fatima, A. L. I. S. Imran, Z. Kastrati, S. M. Daudpota, and A. Soomro, “A Systematic Literature Review on Text Generation Using Deep Neural Network Models,” IEEE Access, vol. 10, pp. 53490–53503, 2022, doi: 10.1109/ACCESS.2022.3174108.

[3] T. Xu, P. Zhang, Q. Huang, and C. V Nov, “AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks”.

[4] S. Reed, Z. Akata, X. Yan, and L. Logeswaran, “Generative Adversarial Text to Image Synthesis,” 2016.

[5] M. Otani, “Toward Verifiable and Reproducible Human Evaluation for Text-to-Image Generation,” pp. 14277–14286.

[6] K. Wang, C. Gou, Y. Duan, Y. Lin, X. Zheng, and F. Y. Wang, "Generative adversarial networks: Introduction and outlook," IEEE/CAA J. Autom. Sin., vol. 4, no. 4, pp. 588–598, 2017, doi: 10.1109/JAS.2017.7510583.

[7] T. Zhang, Z. Wang, J. Huang, M. M. Tasnim, and W. Shi, "A Survey of Diffusion Based Image Generation Models: Issues and Their Solutions," no. 1, 2023, [Online]. Available: http://arxiv.org/abs/2308.13142

[8] B. Ma, Z. Zong, G. Song, H. Li, and Y. Liu, "Exploring the Role of Large Language Models in Prompt Encoding for Diffusion Models," pp. 1–17, 2024, [Online]. Available: http://arxiv.org/abs/2406.11831

[9] M. Zhu, P. Pan, W. Chen, and Y. Yang, "DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2019-June, pp. 5795–5803, 2019, doi: 10.1109/CVPR.2019.00595.

[10] H. Ku and M. Lee, "TextControlGAN: Text-to-Image Synthesis with Controllable Generative Adversarial Networks," Appl. Sci., vol. 13, no. 8, 2023, doi: 10.3390/app13085098.

[11] A. Ramesh et al., "Zero-Shot Text-to-Image Generation," Proc. Mach. Learn. Res. vol. 139, pp. 8821–8831, 2021.

[12] H. Chang et al., "Muse: Text-To-Image Generation via Masked Generative Transformers," Proc. Mach. Learn. Res., vol. 202, pp. 4055–4075, 2023.

[13] W. Wang, Y. Sun, and S. Halgamuge, "Improving MMD-GaN training with repulsive loss function," 7th Int. Conf. Learn. Represent. ICLR 2019, no. February 2019.

[14] X. Jiang, J. Chen, J. Zhang, and Z. Chen, "Towards Text Contextual Understanding: Text Feature Fusion GAN for Text-to-Image Generation Towards Text Contextual Understanding: Text Feature Fusion GAN for Text-to-Image Generation," pp. 0–21, 2024.

[15] C. Saharia et al., "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding," Adv. Neural Inf. Process. Syst., vol. 35, no. NeurIPS, 2022.

[16] M. Liu et al., "LLM4GEN: Leveraging Semantic Representation of LLMs for Text-to-Image Generation," 2024, [Online]. Available: http://arxiv.org/abs/2407.00737

[17] Z. Tan et al., "An Empirical Study and Analysis of Text-to-Image Generation Using Large Language Model-Powered Textual Representation," 2024, [Online]. Available: http://arxiv.org/abs/2405.12914

Disclaimer/Publisher's Note: The perspectives, opinions, and data shared in all publications are the sole responsibility of the individual authors and contributors, and do not necessarily reflect the views of Sciences Force or the editorial team. Sciences Force and the editorial team disclaim any liability for potential harm to individuals or property resulting from the ideas, methods, instructions, or products referenced in the content.