IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 26 January 2024, accepted 6 February 2024, date of publication 9 February 2024, date of current version 20 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3365043

SURVEY

Text-to-Image Synthesis With Generative Models: Methods, Datasets, Performance Metrics, Challenges, and Future Direction

SARAH K. ALHABEEB AND AMAL A. AL-SHARGABI
Department of Information Technology, College of Computer, Qassim University, Buraydah 51452, Saudi Arabia
Corresponding author: Amal A. Al-Shargabi (A.AlShargabi@qu.edu.sa)

This work was supported by the Deanship of Scientific Research, Qassim University.

ABSTRACT Text-to-image synthesis, the process of turning words into images, opens up a world of creative possibilities, and meets the growing need for engaging visual experiences in a world that is becoming more image-based. As machine learning capabilities expanded, the area progressed from simple tools and systems to robust deep learning models that can automatically generate realistic images from textual inputs. Modern, large-scale text-to-image generation models have made significant progress in this direction, producing diversified and high-quality images from text description prompts. Although several methods exist, Generative Adversarial Networks (GANs) have long held a position of prominence. However, diffusion models have recently emerged, with results much beyond those achieved by GANs. This study offers a concise overview of text-to-image generative models by examining the existing body of literature and providing a deeper understanding of this topic. This will be accomplished by providing a concise summary of the development of text-to-image synthesis, previous tools and systems employed in this field, key types of generative models, as well as an exploration of the relevant research conducted on GANs and diffusion models. Additionally, the study provides an overview of common datasets utilized for training the text-to-image model, compares the evaluation metrics used for evaluating the models, and addresses the challenges encountered in the field. Finally, concluding remarks are provided to summarize the findings and implications of the study and open issues for further research.

INDEX TERMS Deep learning, diffusion model, generative models, generative adversarial network, text-to-image synthesis.

I. INTRODUCTION

The rapid improvements made by Artificial Intelligence (AI) in a variety of applications have been remarkable. AI has shown its potential in various ways, and one of the most interesting applications is text-to-image generation. This technology uses natural language processing and computer vision to generate an image based on a given text input. The text provided serves as a set of instructions for the AI's techniques to create an image, which is then rendered in a variety of formats, such as vector graphics, 3D renders, and more. The development of a system that comprehends the relationship between vision and language and can generate visuals that correspond to textual descriptions is a significant step toward achieving an intelligence comparable to that of humans [1].

In recent years, deep learning has allowed for significant progress in the realm of computer vision, allowing for new and improved applications and methods for processing images. Deep learning seeks to discover deep and hierarchical models that accurately describe probability distributions across the many types of data used in AI systems [2]. Image synthesis, the creation of new images and the alteration of old ones, is one such area. Image editing, art generation, computer-aided design, and virtual reality are just a few of the many real-world applications that make image synthesis

The associate editor coordinating the review of this manuscript and approving it for publication was Khursheed Aurangzeb.

an engaging and consequential endeavor [3]. One of the popular approaches is to guide image synthesis with text description, which leads to text-to-image synthesis, which will be addressed in the following section.

A. TEXT-TO-IMAGE SYNTHESIS

Text-to-image synthesis, or the generation of images from text descriptions, is a complex computer vision and machine learning problem that has seen significant progress in recent years. Users may be able to describe visual elements through visually rich text descriptions if automatic image generation from natural language is used. Visual content, like pictures, is a better way to share and understand information because it is more accurate and easy to understand than written text [4].

Text-to-image Synthesis refers to the use of computational methods to convert human-written textual descriptions (sentences or keywords) into visually equivalent representations of those descriptions (images) [3]. The best alignment of visual content matching the text used to be determined through word-to-image correlation analysis combined with supervised methods in synthesis. New unsupervised methods, especially deep generative models, have emerged as a result of recent developments in deep learning. These models are able to generate reasonable visual images by employing appropriately trained neural networks [3]. Figure 1 shows a general architecture of how text-to-image generation would work: a text prompt is fed into an image generative model, which uses the text description to generate an image.

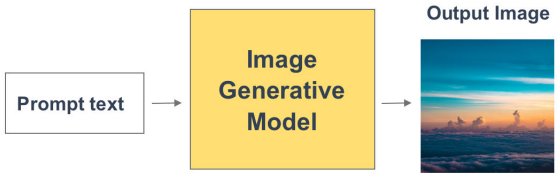

FIGURE 1. General architecture of text-to-image generation.

B. TRADITIONAL METHODS FOR TEXT-TO-IMAGE SYNTHESIS

Early attempts at translating text into images aimed to bridge the gap between humans and machines by emphasizing the importance of natural language comprehension. Some of these systems can take a piece of text written in a natural language and transform it into a sequence of static or dynamic visual representations. Zakraoui et al. [5] conducted an analysis of several established text-to-picture systems and tools, with a focus on identifying challenges and primary issues encountered by previous iterations. This subsection will present an overview of previous text-to-image systems and tools.

The **story picturing engine** [6] applied to the procedure of effectively complementing a narrative with appropriate visual representations. The method is made up of three separate steps: processing the story and choosing photos, figuring out how similar things are, and ranking based on reinforcement. The **text-to-picture synthesis system** [7], aimed to improve communication by generating visual representations based on textual input. The system followed an evolutionary process and adopted semantic role labeling as opposed to keyword extraction, incorporating the concept of picturability to assess the likelihood of identifying a suitable image that represents a given word. To produce compilations of images obtained from the Flickr platform, **Word2Image** [8] implemented a variety of methodologies, including semantic clustering, correlation analysis, and visual clustering.

Moreover, **WordsEye** [9] is a text-to-scene system that mechanically generates static, 3D scenes that are representational of the supplied content. A language analyzer and a visualiser are the two primary parts of the system. Also, a multi-modal system called **CONFUCIUS** [10], which works as a text-to-animation converter, can convert any sentence containing an action verb into an animation that is perfectly synced with speech. A visually assisted instant messaging technique, called **Chat With Illustration (CWI)** [11], automatically provides users with visual messages connected with text messages. Nevertheless, many different systems for other languages exist. In order to handle the Russian language, the **Utkus** [12] text-to-image synthesis system utilizes a natural language analysis module, a stage processing module, and a rendering module. Likewise, **Vishit** [13] is a method for visualizing processed Hindi texts. Language processing, knowledge base construction, and scene generation are its three main computational foundations. Moreover, for the Arabic language, [14] put forth a comprehensive mobile-based system designed for Arabic that generates illustrations for Arabic narratives automatically. The suggested method is specifically designed for utilization on mobile devices, with the aim of instructing Arab children in an engaging and non-traditional manner. Also, using a technique called conceptual graph matching, **Illustrate It!** [15] is a multimedia mobile learning solution for the Arabic language.

C. NEW METHODS FOR TEXT-TO-IMAGE SYNTHESIS

In recent years, scientists have sought a solution to the issue of artificially generating objects. Numerous strategies and technologies have been developed to aid in the generation of new content in various domains, including text, images, audio, etc. Using deep learning approaches, generative models were developed to solve the challenge. The term ‘‘generative modeling’’ describes the process of making fake instances from a dataset that share properties with the original set. The use of generative models makes it possible for machine learning to function with multi-modal outputs [16]. This section demonstrates the four major types of generative models: GAN, Variational Autoencoder, flow-based model, and diffusion model.

1) GENERATIVE ADVERSARIAL NETWORKS

In 2014, Goodfellow et al. [2] introduced GANs, one of the well-known generating models. From that point forward,

several additional models based on the concept of GANs were developed to address the previous shortcomings. GANs can be used in many different contexts, such as to make images of people's faces, to make realistic photos, to make cartoon characters, to age people's faces, to increase resolution, to translate between images and words, and so on [4]. GANs consist of two major sub-models: generator and discriminator. The generator is in charge of making new fake images by taking a noise vector as an input and putting out an image as an output. On the other hand, the discriminator's job is to tell the difference between real and fake images after being trained with real data. In other words, it serves as a classification network that is capable of classifying images by returning 0 for fake and 1 for real. Therefore, the generator's goal is to create convincing fakes in order to trick the discriminator, while the discriminator's goal is to recognize the difference [1]. Training improves both the discriminator's ability to distinguish between real or fake images, and the generator's ability to produce realistic-looking images. When the discriminator can no longer tell genuine images from fraudulent ones, equilibrium has been reached.

2) VARIATIONAL AUTOENCODER (VAE)

The utilization of a variational autoencoder (VAE) [17] provides a probabilistic framework for representing an observation inside a latent space. The input is subjected to encoding, which frequently involves compressing information into a latent space of reduced dimensionality. The primary objective of autoencoders is to effectively encode and represent the given data. The objective at hand involves the identification of a low-dimensional representation for a high-dimensional input, which facilitates the reconstruction of the original input while minimizing the loss of content.

3) FLOW-BASED GENERATIVE MODEL

Flow-based models are capable of learning distinct encoders and decoders. In a manner similar to the encoding phase observed in autoencoders, a transformation is employed to the data, with its parameters determined by a neural network [18]. Nevertheless, the decoder does not consist of a novel neural network that needs to autonomously acquire the decoding process; rather, it functions in direct opposition to its counterpart. In order to achieve the invertibility of a function "f" using neural networks, multiple strategies need to be employed.

4) DIFFUSION MODELS

As a subset of deep generative models, diffusion models have recently been recognized as the cutting edge. The diffusion models have lately demonstrated significant results that have been proven to surpass GAN models [19] They have proven successful in a number of different areas, including the difficult task of image synthesis, where GANs had previously dominated. Recently, diffusion models have become a hot topic in computer vision due to their impressive generative capabilities. The field of generative modeling has found many uses for diffusion models so far, including image generation, super-resolution, inpainting, editing, and translation between images [20]. The principles of non-equilibrium thermodynamics provide the basis for diffusion models. Before learning to rebuild desirable data examples from the noise, they generate a Markov chain of diffusion steps to gradually inject noise into data [20]. In order to learn, the diffusion model has two phases: one for forward diffusion and the other for backward diffusion. In the forward diffusion phase, Gaussian noise is progressively added to the input data at each level [21]. In the second phase, called "reverse," the model is charged to reverse the diffusion process so that the original input data can be recovered.

The architectures of generative model types are shown in Figure 2.

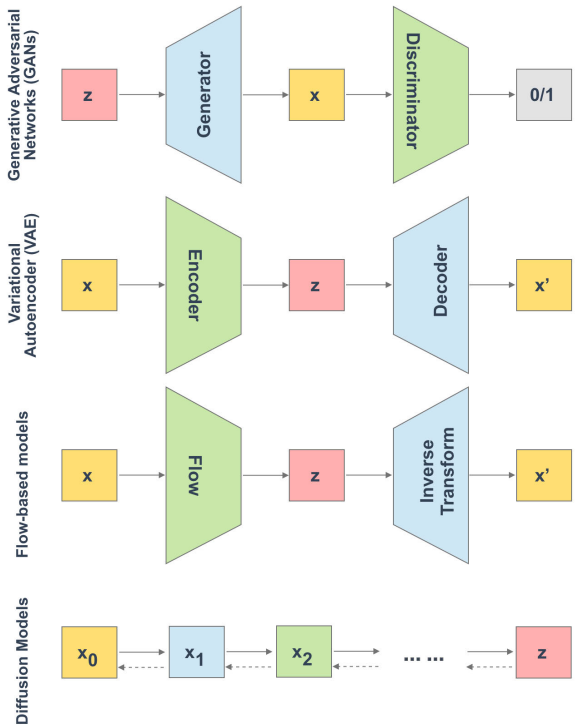

FIGURE 2. Types of generative models, reproduced from Weng [22].

D. RELATED SURVEYS AND STUDY CONTRIBUTION

The state-of-the-art works of GAN-based approaches were examined by Tyagi and Yadav [23], Frolov et al. [1], Zhou et al. [24], and Tan et al. [25]. On the other hand, on the diffusion models field, with multiple works [20], [21], [26], reviewing the progress of diffusion models in all fields, some articles explore deeply into particular areas, including audio diffusion models [27], diffusion models for video generation [28], diffusion model in vision [29], and text-to-image diffusion models [30], providing a thorough overview of the diffusion model field, as well as an in-depth

look at its applications, limitations, and promising future possibilities.

However, our work distinctively integrates the latest advancements in both GANs and diffusion models, providing a holistic view of the field. Unlike the other surveys, which focus primarily on GANs, our review also delves into diffusion models, a cutting-edge area in text-to-image synthesis. Additionally, our paper systematically addresses various research questions, covering a wide array of topics from methods and datasets to evaluation metrics and challenges, offering a broader scope than the previous surveys.

This study focuses on the new approaches to text-to-image synthesis, particularly generative methods, and aims to address five primary questions:

1. RQ1: Which are the existing methods employed, and what are their applications?
2. RQ2: What datasets are commonly used for this purpose?
3. RQ3: What evaluation metrics are used to assess the results?
4. RQ4: What challenges and limitations are associated with the state-of-the-art studies?
5. RQ5: What areas remain unexplored for future research?

II. DATASETS

Datasets play a crucial role in the development and evaluation of text-to-image generative models. In the realm of text-to-image generative models, the utilization of diverse datasets is vital for achieving accurate and realistic visual outputs. This section will explore the various datasets frequently utilized in this research area. The most frequently used datasets by text-to-image synthesis models are:

A. MS COCO

Reference [31], known as the Microsoft Common Objects in Context, is a comprehensive compilation of images that is widely employed for the purpose of object detection and segmentation. The dataset comprises a collection of more than 330,000 images, with each image being accompanied by annotations for 80 object categories and 5 captions that provide descriptive information about the depicted scene. The COCO dataset is extensively utilized in the field of computer vision research and has been employed for the purposes of training and evaluating numerous cutting-edge models for object identification and segmentation.

B. CUB-200-2011

Caltech-UCSD Birds-200-2011 [32] is a popular dataset for fine-grained visual categorization. This dataset comprises 11,788 bird images from 200 subcategories. Images are divided into 5,994 training and 5,794 testing sets. Each image in the dataset has a subcategory, part location, binary attribute, and bounding box labels. Natural language descriptions supplemented these annotations to improve the CUB-200-2011 dataset. Each image received ten single-sentence descriptions.

FIGURE 3. Sample images and their captions of common text-to-image datasets. Figure reproduced from Frolov et al. [1].

C. OXFORD 102 FLOWER

Reference [33] comprises a collection of 102 distinct categories of flowers, which can be effectively employed for image classification. The selected flowers were indigenous to the United Kingdom. The number of photos in each class ranges from 40 to 258. The images demonstrate significant variations in terms of size, pose, and lighting conditions. There exist categories that exhibit significant variations within their respective boundaries, as well as numerous categories that have notable similarities.

Figure 3 shows samples of images along with their captions from the MS COCO, Oxford 102 Flower, and CUB-200-2011 datasets.

D. MULTI-MODAL CELEBA-HQ

A large-scale face image collection, Multi-Modal-CelebA-HQ [34] contains 30,000 high-resolution facial images hand-picked from the CelebA dataset by following CelebA-HQ [35]. Transparent images, sketches, descriptive text, and high-quality segmentation masks accompany each image. Algorithms for face generation and editing, text-guided picture manipulation, sketch-to-image production, and more can all benefit from being trained and tested on the data available in Multi-Modal-CelebA-HQ.

E. CELEBA-DIALOG

Another enormous visual language face dataset with detailed labeling [36], divides a single feature into a range of degrees that all belong to the same semantic meaning. The dataset

has over 200,000 images, encompassing 10,000 distinct identities. Each image is accompanied by five detailed attributes, providing fine-grained information.

F. DEEPFASHION

Reference [37] serves as a valuable resource for the training and evaluating of numerous image synthesis models. It encompasses a comprehensive collection of annotations, including textual descriptions and fine-grained labels, across multiple modalities. The dataset comprises a collection of eight hundred thousand fashion images that exhibit a wide range of diversity, encompassing various accessories and positions.

G. IMAGENET

To test algorithms designed to save, retrieve, or analyze multimedia data, researchers have created a massive database called ImageNet [38], which contains high-quality images that have been manually annotated. There are more than 14 million images in the ImageNet database, all of which have been annotated using the WordNet classification system. Since 2010, the dataset has been applied as a standard for object recognition and image classification in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC).

H. OPENIMAGES

Reference [39] consists of around 9 million images that have been annotated with various types of data, including object bounding boxes, image-level labels, object segmentation masks, localized narratives, and visual relationships. The training dataset of version 7 has 1.9 million images and 16 million bounding boxes representing 600 different item classes, rendering it the most extensive dataset currently available with annotations for object location.

I. CC12M

Conceptual 12M [40] is one of the datasets utilized by OpenAI's DALL-E2 for training, and it consists of 12 million text-image pairs. The dataset, built from the original CC3M dataset of 3 million text-image pairs, was used for a wide range of pre-training and end-to-end training of images.

J. LAION-5B

One of the largest publicly available image-text datasets is Large-scale AI Open Network (LAION) [41]. More than five billion text-image pairs make up LAOIN-5B, an AI training dataset that is 14 times larger than its predecessor, LAOIN-400M.

Table 1 provides a comprehensive comparison of the commonly used datasets used in computer vision and multimodal research. Each dataset is evaluated based on key attributes including domain, common task, number of images, captions per image, training and testing split, and the number of object categories.

III. TEXT-TO-IMAGE GENERATION METHODS

This section provides an overview of relevant studies on text-to-image generative models. Due to the diversity of the generative models and the vast amount of associated literature, this study narrows its focus to the two cutting-edge types of deep learning generative models: GANs and diffusion models.

A. TEXT-TO-IMAGE GENERATION USING GANS

Since its introduction in 2014, GAN-based text-to-image synthesis has been the subject of numerous studies, leading to significant advancements in the field. Reed et al. [42], working upon the foundation laid by deep convolutional GANs [43], were the first to investigate the GAN-based text-to-image synthesis technique.

Earlier models could create images based on universal constraints like a class label or caption, but not pose or location. Therefore, the Generative Adversarial What-Where Network (GAWWN) [44] was proposed, which is a network that generates images based on directions about what to draw and where to draw it. It demonstrates the ability to generate images based on free-form text descriptions and the precise location of objects. GAWWN enables precise location management through the use of a bounding box or a collection of key points.

Stacked Generative Adversarial Networks (StackGAN) [45] established a two-stage conditioning augmentation approach to boost the diversity of synthesized images and stabilize conditional-GAN training. Using the provided text description as input, the Stage-I GAN generates low-resolution images of the initial shape and colors of the object. High-resolution (e.g., 256×256) images with photorealistic features are generated by the Stage-II GAN using the results from Stage-I and the descriptive text.

However, an improvement to this model was made, leading to StackGAN++ [46]. The second version of StackGAN uses generators and discriminators organized in a tree-like structure to produce images at multiple scales that fit the same scene. StackGAN++ has a more reliable training behavior by approximating multiple distributions.

For even more accurate text-to-image production, the Attentional Generative Adversarial Network (AttnGAN) [47] permits attention-driven, multi-stage refining. By focusing on important natural language terms, AttnGAN's attentional-generating network allows it to synthesize fine-grained image features.

To rebuild textual descriptions from the generated images, MirrorGAN [48] presents a text-to-image-to-text architecture with three models. To guarantee worldwide semantic coherence between textual descriptions and the corresponding produced images, it additionally suggests word sentence average embedding.

Figure 4 shows the architectures of: StackGAN, StackGAN++, AttnGAN, and MirrorGAN.

TABLE 1. Overview of commonly used datasets for text-to-image synthesis.

Dataset	Domain	Common Task	Number of Images	Captions per Image	Training Images	Testing Images	Object Categories
MS COCO [31]	General	Object Recognition, Captioning	328K	5	82K	41K	80
CUB-200-2011 [32]	Birds	Fine-grained Classification	11,788K	10	8,855	2,933	200
Oxford 102 Flower [33]	Flowers	Fine-grained Classification	8K	5	7,034	1,155	102
Multi-Modal CelebA-HQ [34]	Celebrities	Multimodal Research, Face Recognition	30K	-	-	-	-
CelebA-Dialog [36]	Celebrities	Face Recognition, Face Image Captioning	200K	6	160K	40K	-
DeepFashion [37]	Fashion	Fashion Recognition, Image Retrieval	800K	-	-	-	50
ImageNet [38]	General	Object Recognition, Classification	14M	-	-	-	21K
OpenImages [38]	General	Object Recognition, Classification	9M	-	-	-	6K
Conceptual 12M [40]	General	Multimodal Research, Captioning	12M	1	-	-	-
LIOAN 5B [41]	General	Multimodal Research, Large-scale Vision	5B	1	-	-	-

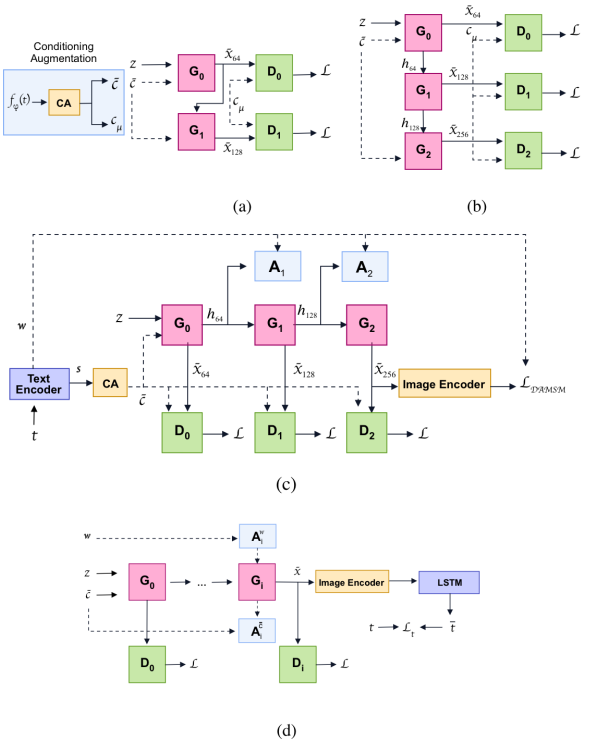

FIGURE 4. The architectures of: (a)StackGAN, (b)StackGAN++, (c)AttnGAN, and (d) MirrorGAN, reproduced from Tan et al. [25].

In the field of story visualization, a story-to-image-sequence generative model, StoryGAN [49], was proposed using the sequential conditional GAN framework. To improve the image resolution and uniformity of the generated sequences, it employs two discriminators, one at the story level and one at the image level, as well as a deep context encoder that dynamically tracks the story flow.

Furthermore, a multi-conditional GAN (MC-GAN) [50] coordinates both the object and the context. The main portion of MC-GAN is a synthesis block that separates object and background information during training. This block helps MC-GAN to construct a realistic object image with the appropriate background by altering the proportion of background and foreground information.

The Dynamic Memory Generative Adversarial Network (DM-GAN) [51] employs a dynamic memory module to enhance the ambiguous image contents in cases where the initial images are generated inadequately. The method can accurately generate images from the text description since a memory writing gate is created to pick the relevant text details based on the content of the initial image. In addition, a response gate is used to adaptively combine the data retrieved from the memories with the attributes of the images.

ManiGAN [52] semantically edits an image to match a provided text describing desirable attributes such as color, texture, and background, while keeping irrelevant content. ManiGAN has two major parts. The first part links visual regions with meaningful phrases for effective manipulation. The second part corrects mismatched properties and completes missing image content.

Without relying on any sort of entanglements between many generators, DeepFusion Generative Adversarial Networks (DF-GAN) [53] may produce high-resolution images directly by a single generator and discriminator. Moreover, DF-GAN's Deep text-image Fusion Block (DFBlock) allows for a more thorough and efficient fusion of text and picture information.

Tedi-GAN [34] combines text-guided image production and modification into one framework for high accessibility, variety, accuracy, and stability in facial image generation and manipulation. It can synthesize high-quality images using multi-modal GAN inversion and a huge multi-modal dataset.

Although there have been many studies on text-to-image generation in English, very few have been applied to other languages. In [54], the use of Attn-GAN was proposed for generating fine-grained images based on descriptions in Bangla text. It is capable of integrating the most exact details in various subregions of the image, with a specific emphasis on the pertinent terms mentioned in the natural language description.

Furthermore, [55] uses language translation models to extend established English text-to-image generating approaches to Hindi text-to-image synthesis. Input Hindi sentences were translated to English by a transformer-based Neural Machine Translation module, whose output was supplied to a GAN-based Image Generation Module.

On the other hand, The CJE-TIG [56] cross-lingual text-to-image pre-training technique removes barriers to using GAN-based text-to-image synthesis models for any given input language. This method alters text-to-image training patterns that are linguistically specific. It uses a bilingual joint encoder in place of a text encoder, applies a discriminator to optimize the encoder, and uses novel generative models to generate content.

The difficulties of visualizing the text of a story with several characters and exemplary semantic relationships were considered in [57]. Two cutting-edge GAN-based image generation models served as inspiration for the researchers' innovative two-stage model architecture for creating images. Stage-I of the image generating process makes use of a scene graph image generation framework; stage-II refines the output image using a StackGAN based on the object layout module and the initial output image. Extensive examination and qualitative results showed that their method could produce a high-quality graphic accurately depicting the text's key concepts.

Short Arabic stories, complete with images that capture the essence of the story and its setting, were offered using a novel approach in [58]. To lessen the need for human input, a text generation method was used in combination with a text-to-image synthesis network. Arabic stories with specialized vocabulary and images were also compiled into a corpus. Applying the approach to the generation of text-image content using various generative models yielded results that proved its value. The method has the potential for use in the classroom to facilitate the development of subject-specific narratives by educators.

A model for generating 256 $\times$ 256 realistic graphics from Arabic text descriptions was proposed in [59]. In order to generate high-quality images, a unique attention network was trained and evaluated in many stages for the proposed model. A deep multimodal similarity model for calculating the loss of matching fine-grained picture text for training the model generator was proposed. The proposed approach set a new standard for converting Arabic text to photorealistic images. On the Caltech-UCSD Birds-200-2011 (CUB) dataset, the newly proposed model produced an inception score of 3.42 $\pm$.05.

Moreover, [60] proposed a robust architecture designed to produce high-resolution realistic images that match a text description written in Arabic. The authors adjusted the shape of the input data to DF-GAN by decreasing the size of the sentence vectors generated by AraBERT. Subsequently, they combined DF-GAN with AraBERT by feeding the sentence embedding vector into the generator and discriminator of DF-GAN. When compared to stackGAN++, their method produced impressive results. In the CUB dataset, it got an FID score of 55.96 and a SI score of 3.51. In the Oxford-102 dataset, got an FID score of 59.45 and a SI score of 3.06.

To improve upon their prior work in [60], the authors presented two additional techniques [61]. To get over the out-of-vocabulary problem, they tried a first technique that involved combining a sample text transformer with the generator and discriminator of DF-GANs. In the second method, the text transformer and training were carried over, and a learning mask predictor was integrated into the architecture to make predictions about masks, which are then utilized as parameters in affine transformations to provide a more seamless fusion between the image and the text. To further improve training stability, the DAMSM loss function was used to train the architecture. The findings proved that the latest technique was superior. Figure 5 shows samples on the CUB dataset, generated by DM-GAN, Attn-GAN, StackGAN, and GAN-INT-CLS.

This study [62] outlines on using transformer-based models (BERT, GPT-2, T5) for text-to-image generation, an under-explored area in computer vision and NLP. It proposes specific architectures to adapt these models for creating images from text descriptions. The study, evaluating the models on challenging datasets, finds that T5 is particularly effective in generating images that are both visually appealing and semantically accurate.

Kang et al. [63] presented a groundbreaking approach to scaling up GANs for text-to-image synthesis. By introducing GigaGAN, a new GAN architecture, the study showcases the ability to generate high-resolution, high-quality images efficiently. GigaGAN demonstrates superior performance in terms of speed and image quality, marking a significant advancement in the use of GANs for large-scale, complex image synthesis tasks.

SWF-GAN, a new model introduced in [64], enhances image synthesis from textual descriptions. It uniquely uses a sentence-word fusion module and a weakly supervised mask predictor for detailed semantic mapping and accurate structure generation. The model effectively creates clear and vivid images with lower computational load, significantly outperforming baseline models in IS and FID scores.

GALIP [65] introduces a novel GAN architecture for text-to-image synthesis. This model integrates transformer-based text encoders and an advanced generator, resulting in high-quality, text-aligned image generation. The model excels in creating images from complex text descriptions, emphasizing the potential of GANs in the realm of text-to-image synthesis.

FIGURE 5. Random image samples on the CUB dataset, generated by DM-GAN, Attn-GAN, StackGAN, and GAN-INT-CLS. Source: [1].

B. TEXT-TO-IMAGE GENERATION USING DIFFUSION MODELS

Unlike GAN-based approaches, which primarily work with small-scale data, autoregressive methods use large-scale data to generate text-to-image conversions, such as DALL-E [66] from OpenAI and Parti [67] from Google. Nevertheless, these approaches have significant computation costs and sequential error buildup due to their autoregressive nature [66], [67], [68], [69]. Conversely, diffusion models are highly popular for all sorts of generating applications.

To create images from text, the study [70] introduced the vector quantized diffusion (VQ-Diffusion) model. Vector quantized variational autoencoders (VQ-VAEs) form the basis of this technique, with the latent space being modeled using a conditional variant of the Denoising Diffusion Probabilistic Model (DDPM). Using a natural language description with an ROI mask, the Blended Diffusion approach was provided in [71] for making local (region-based) adjustments to real images. The authors were successful in their mission by employing a pretrained language-image model (CLIP) to guide the modification in the direction of a given text prompt and combining it with a DDPM to generate results that looked natural.

CLIP-Forge [72] was proposed as a solution to the widespread absence of coupled text and shape data. Utilizing a two-step training approach, CLIP-Forge requires only a pre-trained image-text network like CLIP, as well as an unlabeled shape dataset. One of the advantages of this approach is that it can produce various shapes for a given text without resorting to costly inference time optimization.

In [73], the authors investigate CLIP guidance and classifier-free guidance as two separate guiding methodologies for the problem of text-conditional image synthesis. Their proposed model, GLIDE, which stands for Guided Language to Image Diffusion for Generation and Editing, was shown to be the most liked by humans in terms of caption similarity and photorealism. It also often made examples that were very photorealistic.

Specifically for conditional image synthesis, M6-UFC was presented in [74] as a universal form for unifying several multi-modal controls. To quicken inference, boost global consistency, and back up preservation controls, the authors turned to non-autoregressive generation. In addition, they developed a progressive generation process using relevance and fidelity estimators to guarantee accuracy.

Using the language-image priors retrieved from a pre-trained CLIP model, this study [75] proposes a self-supervised approach called CLIP-GEN for automatic text-to-image synthesis. Here, a text-to-image generator can be taught to work with just a group of images from the broad domain that don't have labels. This will help to prevent the need to collect vast amounts of matched text-image data, which is too costly to gather.

Imagen, a method for text-to-image synthesis presented in [76], uses a single encoder for the text sequence and a set of diffusion models to generate high-resolution images. The text embeddings provided by the encoder are also a prerequisite for these models. As an added bonus, the authors presented a brand new caption set (DrawBench) for testing text-to-image conversion. The authors created Efficient U-Net, an efficient network architecture, and used it in their text-to-image generation experiments to test its efficacy. Figure 6 represents a simple visualisation of Imagen architecture.

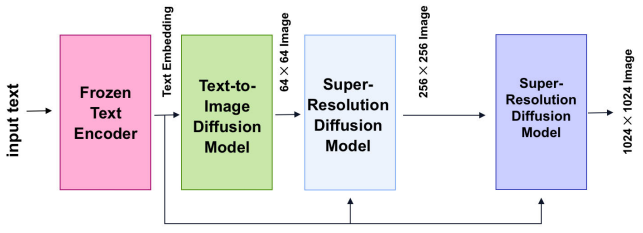

FIGURE 6. Overview of Imagen, reproduced from Saharia et al. [76].

ERNIE-ViLG 2.0 [77] is a large-scale Chinese text-to-image diffusion model that uses fine-grained textual and visual information about important parts of the scene along

TABLE 2. Diffusion Models-based related studies.

Ref.	Year	Model	Dataset
[70]	2021	VQ-Diffusion	CUB-200, Oxford-102 & MS-COCO
[73]	2021	GLIDE	250M image-text pairs
[79]	2022	Stable Diffusion	LAION dataset
[80]	2022	DALL-E-2	650M images
[74]	2021	M6-UFC	M2C-Fashion & Multi-Modal CelebA-HQ
[75]	2022	CLIP-GEN	MS-COCO & ImageNet
[76]	2022	Imagen	860M text-image pairs
[77]	2022	ERNIE-ViLG 2.0	170M image-text pair
[78]	2022	eDiff-I	1B text-image pairs
[81]	2022	DiVAE	ImageNet

with different denoising specialists at different denoising stages to improve the quality of the output images.

On the other hand, eDiff-I [78] outperforms other large-scale text-to-image diffusion models by improving text alignment while keeping inference computation cost and visual quality stable. Unlike traditional diffusion models, which rely on a single model trained to denoise the entire noise distribution, eDiff-I is instead trained on an ensemble of expert denoisers, each of which is tailored to denoising at a distinct stage of generation. The researchers claim that employing such specialized denoisers enhances the quality of synthesized output.

Frido [82] is an image-synthesizing Feature Pyramid Diffusion model that conducts multiscale coarse-to-fine denoising. To construct an output image, it first decomposes the input into vector quantized scale-dependent components. The previously mentioned stage of learning multi-scale representations can also take advantage of input conditions such as language, scene graphs, and image layout. Frido can thus be utilized for both traditional and cross-modal image synthesis.

A new method called DreamBooth was suggested in [83] as a way to tailor the results of text-to-image generation from diffusion models to the needs of users. The authors fine-tuned a pretrained text-to-image model so that it is able to associate a distinctive identifier with a subject given only a small number of images of that subject as input. Following the subject's incorporation into the model's output domain, the identifier can be used to generate completely brand-new photorealistic pictures of the subject in a variety of settings.

Furthermore, Imagic [84] shows how a single real image can be subjected to sophisticated text-guided semantic edits. While maintaining the image's original qualities, Imagic can alter the position and composition of one or more objects within it. It works on raw images without the need for image masks or any other preprocessing.

Likewise, UniTune [85] is capable of editing images with a high degree of semantic and visual fidelity to the original, given a random image and a textual edit description as input. It can be considered an art-direction tool that only requires text as input rather than more complex requirements such as masks or drawings.

DiVAE, a VQ-VAE architecture model that employs a diffusion decoder as the reconstructing component in image synthesis, was proposed by Shi et al. in [81]. They investigated how to incorporate image embedding into the diffusion model for high performance and discovered that a minor adjustment to the U-Net used in diffusion could accomplish this.

Building upon the success of its predecessor [66], DALL-E 2 [80] was launched as a follow-up version with the intention of producing more realistic images at greater resolutions by combining concepts, features, and styles. The model consists of two parts: a prior that creates a CLIP image embedding from a caption and a decoder that creates an image based on the embedding. It was demonstrated that increasing image variety through the intentional generation of representations leads to only a slight decrease in photorealism and caption similarity.

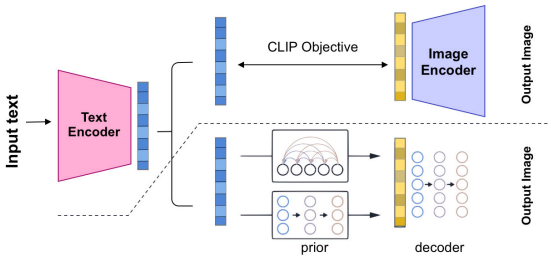

FIGURE 7. Overview of DALL-E 2, reproduced from Ramesh et al. [80].

Figure 7 represents an overview of DALL-E 2, and Figure 8 shows samples of images generated by DALL-E 2 given a detailed text prompt.

Furthermore, the advanced model DALL-E 3 [86], which was recently released, represents a significant advancement over its predecessors. Leveraging advanced diffusion models, DALL-E 3 not only excels in maintaining fidelity to textual prompts but also underscores its ability to capture intricate details, marking a substantial advancement in the realm of generative models.

Stable Diffusion is another popular text-to-image tool that was introduced in 2022, based on a previous work [79]. Stable Diffusion employs a type of diffusion model known as the latent diffusion model (LDM). The VAE, U-Net, and an optional text encoder comprise Stable Diffusion. Compared to pixel-based diffusion models, LDMs dramatically reduced the requirement for processing while achieving a new state-of-the-art picture inpainting and highly competitive performance on a variety of applications like unconditional image creation and super-resolution. Figure 9 shows an overview of the architecture of Stable diffusion.

Table 2 summarizes the studies that utilized diffusion models in text-to-image generation by year, model, and dataset.

FIGURE 8. Samples generated by DALL-E 2 given the prompt: "a bowl of soup that is a portal to another dimension as digital art". Source: [80].

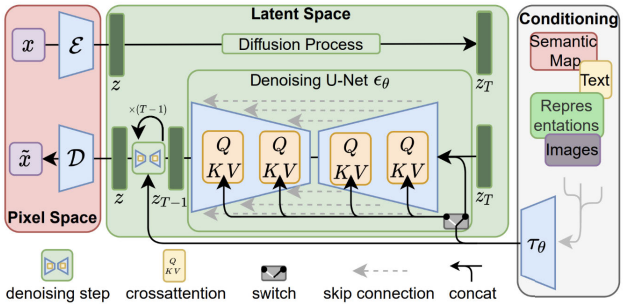

FIGURE 9. Overview of stable diffusion. Source: [79].

ParaDiffusion [87] is an innovative text-to-image generation model adept at transforming detailed, long-form text into corresponding images. It stands out due to its deep semantic understanding derived from large language models, enabling it to create images that are both visually appealing and closely aligned with complex textual descriptions. The model's training is enhanced by the ParaImage Dataset, which includes extensive image-text pairs. This approach marks a significant advancement in AI-driven media, particularly in generating intricate images from elaborate text descriptions.

UPainting is an approach that was presented in [88] to automatic painting generation using deep learning. The model captures the essence of famous painters and styles, enabling the creation of new artworks that reflect the characteristics of these styles. It's a blend of art and technology, offering a new way of creating art with AI's assistance.

CLIPAG [89] explores a unique approach to text-to-image generation without relying on traditional generative models. It leverages Perceptually Aligned Gradients (PAG) in robust Vision-Language models, specifically an enhanced version of CLIP, to generate images directly aligned with text descriptions. This method marks a shift in text-to-image synthesis, utilizing a more streamlined and efficient process compared to conventional methods.

GLIGEN was proposed in [90] as a new method for text-to-image generation, focusing on generating linguistically coherent and visually compelling images. It emphasizes the integration of natural language understanding and image synthesis, demonstrating impressive capabilities in creating images that accurately reflect complex textual inputs.

Snapfusion [91] introduces an efficient text-to-image diffusion model optimized for mobile devices, achieving image generation in under two seconds. It addresses the computational intensity and speed limitations of existing diffusion models through an innovative network architecture and improved step distillation. The proposed UNet efficiently synthesizes high-quality images, outperforming the baseline Stable Diffusion model in terms of FID and CLIP scores.

Zhang et al. [92] introduced a method to add conditional control to image generation models, allowing for more precise and tailored image creation. The approach improves the ability to generate images that meet specific criteria or conditions, enhancing the versatility and applicability of image-generation technologies.

Moreover, Zhao et al. [93] explored advancements in text-to-image diffusion models, focusing on enhancing their capabilities to produce more realistic and varied images. The study delves into new methods and techniques to improve these models, significantly advancing the field of T2I synthesis.

The researchers in [94] focused on adapting the English Stable Diffusion model for Chinese text-to-image synthesis. They introduced a novel method for transferring the model's capabilities to the Chinese language, resulting in high-quality image generation from Chinese text prompts, significantly reducing the need for extensive training data.

AltDiffusion [95] presents a multilingual text-to-image diffusion model supporting eighteen languages, addressing the limitations of existing models that cater primarily to English. The paper details the development and effectiveness of this model in generating culturally relevant and accurate images across various languages, showcasing its potential for global use in T2I tasks.

Random image samples on the MS-COCO dataset are represented in Figure 10, generated by DALL-E, GLIDE, and DALL-E 2.

IV. EVALUATION METRICS

The majority of current metrics evaluate a model's quality by considering two main factors: the quality of the images it produces and the alignment between text and images. Fréchet Inception Distance (FID) [96] and Inception Score (IS) [97] are commonly used metrics for appraising the image quality of a model. These metrics were initially developed for traditional GAN tasks focused on assessing image quality. To evaluate text-image alignment, the R-precision [47] metric is widely employed.

FIGURE 10. Random image samples on MS-COCO, generated by DALL-E, GLIDE, and DALL-E 2. Source: [80].

For more in-depth details, we refer to [98]. Moreover, the Clip Score [99] is used in evaluating common sense and mentioned objects, while Human Evaluation offers a comprehensive insight into multiple aspects of image generation. In the following a detailed description of each metric.

A. THE FRECHET INCEPTION DISTANCE (FID) [96]

Using the feature space of a pre-trained Inception v3 network, FID [77] determines the frechet distance between natural and artificial distributions. This equation solves it:

$$F(r,g) = \left\|\mu_r - \mu_g\right\|^2 + trace\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r\Sigma_g\right)^{\frac{1}{2}}\right) \tag{1}$$

where r and g represent, respectively, the image's real and generated features. The covariance and mean of real and produced features are represented by r, g, r, and g, correspondingly. The lower FID score is considered to be the more appropriate score. It describes the level of realism, accuracy, and variety in the generated distributions. Table 3 represents a comparison of FID scores obtained by GANs and diffusion models on the MS-COCO dataset and shows that diffusion models made remarkable results.

B. THE INCEPTION SCORE

Reference [97], which ignores the underlying distribution, measures the produced distribution's faithfulness and diversity. The following is the IS equation:

$$I = exp(\mathbb{E}_x D_{KL}(p(y|x)\|p(y))) \tag{2}$$

IS calculates the difference between the marginal distribution p(y) and the conditional distribution p(y|x) using the Kull back-Leibler (KL) divergence. The generated image x, denoted by the label y, is predicted using a pre-trained Inception v3 network. Unlike FID, a higher IS is preferable. It implies high-quality images accurately categorized by class.

TABLE 3. FID scores of GANs and diffusion models on the MS-COCO dataset.

Ref.	Model	FID ↓
[45]	Stackgan	74.05
[46]	Stackgan++	81.59
[47]	AttnGAN	35.49
[48]	MirrorGAN	-
[51]	DM-GAN	32.64
[53]	DF-GAN	21.42
[70]	VQ-Diffusion	13.86
[73]	GLIDE	12.24
[79]	Stable Diffusion	12.63
[80]	DALL-E-2	10.39
[76]	Imagen	7.27
[77]	ERNIE-ViLG 2.0	6.75
[78]	eDiff-I	6.95

C. THE R-PRECISION (RP)

Reference [47] metric is widely employed for assessing the consistency between text and images. RP operates on the principle of employing a generated image query based on the provided caption. Specifically, given an authentic text description and 99 randomly selected mismatched captions, an image is produced from the authentic caption. This resulting image is then utilized to query the original description from a pool of 100 candidate captions. The retrieval is deemed successful if the similarity score between it and the authentic caption is the highest. The matching score is determined using the cosine similarity between the encoding vectors of the image and the caption. A higher RP score indicates better quality, with RP being the proportion of successful retrievals.

D. CLIP SCORE

The CLIP model [99], developed by OpenAI, demonstrates the ability to evaluate the semantic similarity between a given text caption and an accompanying image. Based on this rationale, the CLIP score can serve as a quantitative measure and is formally defined as:

$$\mathbb{E}[s(f(image) * g(caption))] \tag{3}$$

where the mathematical expectation is computed over the set of created images in a batch, and s represents the logarithmic scale of the CLIP logit [73]. A higher Clip score suggests a stronger semantic relationship between the image and the text, while a lower score shows less of a connection.

E. HUMAN EVALUATIONS

Some studies used human evaluation as a qualitative measure to assess and evaluate the quality of the results. The reporting of metrics based on human evaluation is motivated by the fact that many possible applications of the models are centered upon tricking the human observer [100]. Typically, a collection of images is provided to an individual, who is tasked with evaluating their quality in terms of photorealism and alignment with associated captions.

Frolov et al. [1] proposed a set of different criteria for comparing evaluation metrics. The following is an explanation of these criteria.

- **Image Quality and Diversity**: The degree to which the generated image looks realistic or similar to the reference image and the ability of the model to produce varied images based on the same text prompt.
- **Text Relevance**: How well the generated image corresponds to the given text prompt.
- **Mentioned Objects and Object Fidelity**: Whether the model correctly identifies and includes the objects mentioned in the text, and how accurately the objects in the generated image match their real-world counterparts.
- **Numerical and Positional Alignment**: The accuracy of any quantitative details and the positional arrangement of objects in the generated image in relation to the provided text.
- **Common Sense**: The presence of logical and expected elements in the generated image.
- **Paraphrase Robustness**: The model remains unaffected by minor modifications in the input description, such as word substitutions or rephrasings.
- **Explainable**: The ability to provide a clear explanation of why an image is not aligned with the input.
- **Automatic**: Whether the metric can be calculated automatically without human intervention.

Based on these Key criteria, we provide in Table 4 a comparative analysis of the commonly used text-to-image evaluation metrics based on their performance. It is important to note that the table presented offers a simplified overview. In practice, choosing the right metric depends on the specific goals and context of the text-to-image generation task. Additionally, the effectiveness of these metrics may vary depending on the specific model and dataset used.

V. CHALLENGES AND LIMITATIONS

Although there has been significant progress made in the area of creating visual representations of textual descriptions, there are still some challenges and limitations that will be discussed below.

A. OPEN SOURCE

Although DALL-E is one of the competitive models, unfortunately, it has not been released for public usage. There is a copy of DALL-E 2 available in PyTorch [101], but no pre-trained model. However, the Stable Diffusion model is among the open-source models that are currently accessible. Stable Diffusion benefits from extensive community support due to its open-source nature. Consequently, it is anticipated that there will be additional advancements in this particular area in the near future.

B. LANGUAGE SUPPORT

The majority of studies in the field of text-to-image generation have been conducted on English text descriptions due to the abundance of dataset resources and the simple structure of

TABLE 4. Overview of commonly used evaluation metrics for text-to-image synthesis, adapted from Frolov et al. [1].

Metric	Image Quality	Image Diversity	Text Relevance	Mentioned Objects	Object Fidelity	Numerical Alignment	Positional Alignment	Common Sense	Paraphrase Robustness	Explainable	Automatic
FID	✓	✓									✓
IS	✓										✓
R-precision (RP)			✓								✓
Clip Score			✓					✓			✓
Human Evaluation	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	

the language. Some languages, however, require more effort which needs to be addressed. For instance, Arabic, in contrast to English, has more complicated morphological features and fewer semantic and linguistic resources [5]. This is a main challenge that needs to be dealt with in text-to-image generation.

C. COMPUTATIONAL COMPLEXITY

The computational complexity of diffusion models poses a notable difficulty. The process of training a diffusion model involves multiple iterative processes which can impose a significant computational burden. Therefore, The model's scalability may be constrained by the increased complexity observed when working with larger datasets and higher-resolution images. Moreover, for further research in the field of text-to-image generative models, and despite the availability of big datasets like LION-5B to the general public, the utilization of such datasets remains challenging for individuals due to the substantial hardware requirements involved.

D. ETHICAL CONSIDERATIONS

It is important to consider the potential ethical issues that arise with the use of text-to-image generative models. One of the significant concerns is the potential for misuse of these models. With the ability to generate realistic images based on text descriptions, there is a risk that these models could be used to create deceptive or misleading content. This could have serious consequences in various areas, such as fake news, fraud, or even harassment.

Another issue is the potential bias that can be embedded in the generated images. If the training data used to develop these models is not diverse and representative, there is a possibility that the generated images may reflect prejudices or stereotypes present in the data.

VI. FUTURE DIRECTIONS

The domain of text-to-image generation is experiencing significant advancements on a regular basis. The recent emergence of novel generative diffusion models, including DALL-E, Midjourney, Stable diffusion, and others, has sparked significant interest and discussion in the scientific community. The field shows a high degree of fertility and renewability, as seen by the recent publication of numerous relevant studies and an ongoing flow of new papers within a relatively short timeframe.

By making generative models open-source, researchers and developers can collaborate more effectively, which will in turn boost innovation in the field. Researchers may utilize these publicly available models to investigate novel uses, enhance current AI models, and move the field forward rapidly.

To overcome the language barrier, some studies proposed multilingual [95] and cross-lingual [56] models to support multiple languages within the same model. The goal of these multilingual models is to break down linguistic barriers by providing a common groundwork for the comprehension and processing of several languages at once. This method has the ability to dramatically improve linguistic diversity in communication and open up access to information for everyone.

Moreover, to make these technologies more widely accessible and sustainable, it will be essential to improve resource efficiency and minimize computational complexity by creating models that produce high-quality photos using fewer computer resources.

Nevertheless, greater research into ethical and bias considerations is required. Ensuring fairness, removing bias, and following ethical rules are still critical considerations for any AI system. Possible directions for future study in this area include developing models with increased consciousness and sensitivity to these factors.

The utilization of text-to-image production exhibits a wide range of applications across several domains, including but not limited to education, product design, and marketing. This technology enables the creation of visual materials, such as illustrations and infographics, that seamlessly integrate text and images. There are some early assumptions about which businesses might be impacted by the growing area of image generation, which will have an impact on any sector that relies on visual art, such as graphic design, filmmaking, or photography [100].

VII. CONCLUSION

The field of text-to-image synthesis has made significant progress in recent years. The development of GANs and diffusion models has paved the way for more advanced and realistic image generation from textual descriptions. These models have demonstrated an outstanding ability to generate high-quality images across a wide range of domains and datasets. This study offers a comprehensive review of the existing literature on text-to-image generative models, summarizing the historical development, popular datasets, key methods, commonly used evaluation metrics, and challenges faced in this field. Despite these challenges, the potential of text-to-image generation in expanding creative horizons and enhancing AI systems is undeniable. The ability to generate realistic and diverse images from textual inputs opens up new possibilities in various fields, including art, design, advertising, and others. Therefore, researchers and practitioners should continue to explore and refine text-to-image generative models.

ACKNOWLEDGMENT

The authors would like to thank the Deanship of Scientific Research, Qassim University, for funding the publication of this project.

REFERENCES

[1] S. Frolov, T. Hinz, F. Raue, J. Hees, and A. Dengel, "Adversarial text-to-image synthesis: A review," *Neural Netw.*, vol. 144, pp. 187–209, Dec. 2021.

[2] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014, *arXiv:1406.2661*.

[3] J. Agnese, J. Herrera, H. Tao, and X. Zhu, "A survey and taxonomy of adversarial neural networks for text-to-image synthesis," *WIREs Data Mining Knowl. Discovery*, vol. 10, no. 4, Jul. 2020, Art. no. e1345.

[4] L. Jin, F. Tan, and S. Jiang, "Generative adversarial network technologies and applications in computer vision," *Comput. Intell. Neurosci.*, vol. 2020, pp. 1–17, Aug. 2020.

[5] J. Zakraoui, M. Saleh, and J. A. Ja'am, "Text-to-picture tools, systems, and approaches: A survey," *Multimedia Tools Appl.*, vol. 78, no. 16, pp. 22833–22859, Aug. 2019, doi: 10.1007/s11042-019-7541-4.

[6] D. Joshi, J. Z. Wang, and J. Li, "The story picturing engine—A system for automatic text illustration," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 2, no. 1, pp. 68–89, Feb. 2006, doi: 10.1145/1126004.1126008.

[7] X. Zhu, A. Goldberg, M. Eldawy, C. Dyer, and B. Strock, "A text-to-picture synthesis system for augmenting communication," in *Proc. 22nd AAAI Conf. Artif. Intell.*, 2007, p. 1590.

[8] H. Li, J. Tang, G. Li, and T.-S. Chua, "Word2Image: Towards visual interpreting of words," in *Proc. 16th ACM Int. Conf. Multimedia*, 2008, pp. 813–816.

[9] B. Coyne and R. Sproat, "WordsEye: An automatic text-to-scene conversion system," in *Proc. 28th Annu. Conf. Comput. Graph. Interact. Techn.*, Aug. 2001, pp. 487–496.

[10] M. E. Ma, "Confucius: An intelligent multimedia storytelling interpretation and presentation system," School Comput. Intell. Syst., Univ. Ulster, Coleraine, U.K., Tech. Rep., 2002.

[11] Y. Jiang, J. Liu, and H. Lu, "Chat with illustration," *Multimedia Syst.*, vol. 22, no. 1, pp. 5–16, Feb. 2016, doi: 10.1007/s00530-014-0371-3.

[12] D. Ustalov, "A text-to-picture system for Russian language," in *Proc. 6th Russian Young Scientists Conf. Inf. Retr.*, Aug. 2012, pp. 35–44.

[13] P. Jain, H. Darbari, and V. C. Bhavsar, "Vishit: A visualizer for Hindi text," in *Proc. 4th Int. Conf. Commun. Syst. Netw. Technol.*, Apr. 2014, pp. 886–890.

[14] A. G. Karkar, J. M. Al Ja'am, S. Foufou, and A. Sleptchenko, "An e-learning mobile system to generate illustrations for Arabic text," in *Proc. IEEE Global Eng. Educ. Conf.*, Apr. 2016, pp. 184–191.

[15] A. G. Karkar, J. M. Alja'am, and A. Mahmood, "Illustrate it! An Arabic multimedia text-to-picture m-learning system," *IEEE Access*, vol. 5, pp. 12777–12787, 2017.

[16] I. Goodfellow, "NIPS 2016 tutorial: Generative adversarial networks," 2017, *arXiv:1701.00160*.

[17] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," *Found. Trends Mach. Learn.*, vol. 12, no. 4, pp. 307–392, 2019.

[18] L. Weng. (2018). *Flow-based Deep Generative Models*. [Online]. Available: https://lilianweng.github.io/posts/2018-10-13-flow-models/

[19] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," 2021, *arXiv:2105.05233*.

[20] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," 2022, *arXiv:2209.04747*.

[21] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, Y. Shao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *Comprehensive Surv. Methods Appl.*, vol. 1, p. 39, Sep. 2022.

[22] L. Weng. (2021). *What Are Diffusion Models*. [Online]. Available: https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

[23] S. Tyagi and D. Yadav, "A comprehensive review on image synthesis with adversarial networks: Theory, literature, and applications," *Arch. Comput. Methods Eng.*, vol. 29, no. 5, pp. 2685–2705, Aug. 2022.

[24] R. Zhou, C. Jiang, and Q. Xu, "A survey on generative adversarial network-based text-to-image synthesis," *Neurocomputing*, vol. 451, pp. 316–336, Sep. 2021.

[25] Y. X. Tan, C. P. Lee, M. Neo, K. M. Lim, J. Y. Lim, and A. Alqahtani, "Recent advances in text-to-image synthesis: Approaches, datasets and future research prospects," *IEEE Access*, vol. 11, pp. 88099–88115, 2023.

[26] H. Cao, C. Tan, Z. Gao, Y. Xu, G. Chen, P.-A. Heng, and S. Z. Li, "A survey on generative diffusion model," 2022, *arXiv:2209.02646*.

[27] C. Zhang, C. Zhang, S. Zheng, M. Zhang, M. Qamar, S.-H. Bae, and I. S. Kweon, "A survey on audio diffusion models: Text to speech synthesis and enhancement in generative AI," 2023, *arXiv:2303.13336v2*.

[28] R. Yang, P. Srivastava, and S. Mandt, "Diffusion probabilistic modeling for video generation," 2022, *arXiv:2203.09481v5*.

[29] A. Ulhaq, N. Akhtar, and G. Pogrebna, "Efficient diffusion models for vision: A survey," 2022, *arXiv:2210.09292v2*.

[30] C. Zhang, C. Zhang, M. Zhang, and I. So Kweon, "Text-to-image diffusion models in generative AI: A survey," 2023, *arXiv:2303.07909v2*.

[31] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Computer Vision—ECCV 2014* (Lecture Notes in Computer Science). Cham, Switzerland: Springer, 2014, pp. 740–755.

[32] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "Caltech-UCSD birds 200," California Inst. Technol., Tech. Rep. CNS-TR-2011-001, 2011.

[33] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. 6th Indian Conf. Comput. Vis., Graph. Image Process.*, Dec. 2008, pp. 722–729.

[34] W. Xia, Y. Yang, J.-H. Xue, and B. Wu, "TediGAN: Text-guided diverse face image generation and manipulation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Dec. 2020, pp. 2256–2265.

[35] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. 6th Int. Conf. Learn. Represent.*, Oct. 2018.

[36] Y. Jiang, Z. Huang, X. Pan, C. C. Loy, and Z. Liu, "Talk-to-edit: Fine-grained facial editing via dialog," 2021, *arXiv:2109.04425*.

[37] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "DeepFashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1096–1104.

[38] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255. [Online]. Available: http://www.image-net.org

[39] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, T. Duerig, and V. Ferrari, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *Int. J. Comput. Vis.*, vol. 128, no. 7, pp. 1956–1981, Jul. 2020.

[40] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3557–3567.
[41] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev, "LAION-5B: An open large-scale dataset for training next generation image-text models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022.
[42] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," 2016, *arXiv:1605.05396*.
[43] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*.
[44] S. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, "Learning what and where to draw," 2016, *arXiv:1610.02454*.
[45] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5908–5916.
[46] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "StackGAN++: Realistic image synthesis with stacked generative adversarial networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1947–1962, Aug. 2019. [Online]. Available: https://github.com/hanzhanggit/StackGAN-v2.
[47] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1316–1324.
[48] T. Qiao, J. Zhang, D. Xu, and D. Tao, "MirrorGAN: Learning text-to-image generation by redescription," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1505–1514.
[49] Y. Li, Z. Gan, Y. Shen, J. Liu, Y. Cheng, Y. Wu, L. Carin, D. Carlson, and J. Gao, "StoryGAN: A sequential conditional GAN for story visualization," 2018, *arXiv:1812.02784*.
[50] H. Park, Y. Yoo, N. K. Mc-Gan, H. Park, Y. Yoo, and N. Kwak, "MC-GAN: Multi-conditional generative adversarial network for image synthesis," in *Proc. Brit. Mach. Vis. Conf.*, May 2018, p. 76.
[51] M. Zhu, P. Pan, W. Chen, and Y. Yang, "DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5795–5803.
[52] B. Li, X. Qi, T. Lukasiewicz, and P. H. S. Torr, "ManiGAN: Text-guided image manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7877–7886.
[53] M. Tao, H. Tang, F. Wu, X. Jing, B.-K. Bao, and C. Xu, "DF-GAN: A simple and effective baseline for text-to-image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16494–16504.
[54] M. A. Haque Palash, M. A. Al Nasim, A. Dhali, and F. Afrin, "Fine-grained image generation from Bangla text description using attentional generative adversarial network," in *Proc. IEEE Int. Conf. Robot., Autom., Artif.-Intell. Internet-of-Things (RAAICON)*, Dec. 2021, pp. 79–84. [Online]. Available: https://ieeexplore.ieee.org/document/9929536/
[55] A. S. Parihar, A. Kaushik, A. V. Choudhary, and A. K. Singh, "HTGAN: An architecture for Hindi text based image synthesis," in *Proc. 5th Int. Conf. Comput., Commun. Signal Process. (ICCCSP)*, May 2021, pp. 273–279.
[56] H. Zhang, S. Yang, and H. Zhu, "CJE-TIG: Zero-shot cross-lingual text-to-image generation by corpora-based joint encoding," *Knowl.-Based Syst.*, vol. 239, Mar. 2022, Art. no. 108006.
[57] J. Zakraoui, M. Saleh, S. Al-Maadeed, and J. M. Jaam, "Improving text-to-image generation with object layout guidance," *Multimedia Tools Appl.*, vol. 80, no. 18, pp. 27423–27443, Jul. 2021, doi: 10.1007/s11042-021-11038-0.
[58] J. Zakraoui, S. A. Maadeed, M. S. A. El-Seoud, J. M. Alja'am, and M. Salah, "A generative approach to enrich Arabic story text with visual aids," in *Proc. 10th Int. Conf. Softw. Inf. Eng.* New York, NY, USA: Association for Computing Machinery, 2021, pp. 47–52, doi: 10.1145/3512716.3512725.
[59] S. M. Mathematics and M. Loey, "Photo realistic generation from Arabic text description based on generative adversarial networks," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, Mar. 2022, doi: 10.1145/3490504.
[60] M. Bahani, A. El Ouaazizi, and K. Maalmi, "AraBERT and DF-GAN fusion for Arabic text-to-image generation," *Array*, vol. 16, Dec. 2022, Art. no. 100260. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2590005622000935
[61] M. Bahani, S. M. Ben, K. Maalmi, and A. E. Ouaazizi. (Oct. 2022). *Increase the Effectiveness of the Arabic Text-to-image Generation Task*. [Online]. Available: https://www.researchsquare.com/article/rs-2169841/v1
[62] M. Bahani, A. E. Ouaazizi, and K. Maalmi, "The effectiveness of T5, GPT-2, and BERT on text-to-image generation task," *Pattern Recognit. Lett.*, vol. 173, pp. 57–63, Sep. 2023.
[63] M. Kang, J.-Y. Zhu, R. Zhang, J. Park, E. Shechtman, S. Paris, and T. Park, "Scaling up GANs for text-to-image synthesis," 2023, *arXiv:2303.05511v2*.
[64] C. Liu, J. Hu, and H. Lin, "SWF-GAN: A text-to-image model based on sentence-word fusion perception," *Comput. Graph.*, vol. 115, pp. 500–510, Oct. 2023.
[65] M. Tao, B.-K. Bao, H. Tang, and C. Xu, "GALIP: Generative adversarial CLIPs for text-to-image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14214–14223.
[66] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," 2021, *arXiv:2102.12092v2*.
[67] J. Yu, Y. Xu, J. Yu Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B. K. Ayan, B. Hutchinson, W. Han, Z. Parekh, X. Li, H. Zhang, J. Baldridge, and Y. Wu, "Scaling autoregressive models for content-rich text-to-image generation," 2022, *arXiv:2206.10789v1*.
[68] M. Ding, Z. Yang, W. Hong, W. Zheng, C. Zhou, D. Yin, J. Lin, X. Zou, Z. Shao, H. Yang, and J. Tang, "CogView: Mastering text-to-image generation via transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 24, May 2021, pp. 19822–19835.
[69] M. Ding, W. Zheng, W. Hong, and J. Tang, "CogView2: Faster and better text-to-image generation via hierarchical transformers," 2022, *arXiv:2204.14217v2*.
[70] S. Gu, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, L. Yuan, and B. Guo, "Vector quantized diffusion model for text-to-image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nov. 2021, pp. 10686–10696.
[71] O. Avrahami, D. Lischinski, and O. Fried, "Blended diffusion for text-driven editing of natural images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18187–18197.
[72] A. Sanghi, H. Chu, J. G. Lambourne, Y. Wang, C.-Y. Cheng, M. Fumero, and K. R. Malekshan, "CLIP-Forge: Towards zero-shot text-to-shape generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Oct. 2021, pp. 18582–18592.
[73] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models," 2021, *arXiv:2112.10741v3*.
[74] Z. Zhang, J. Ma, C. Zhou, R. Men, Z. Li, M. Ding, J. Tang, J. Zhou, and H. Yang, "M6-UFC: Unifying multi-modal controls for conditional image synthesis via non-autoregressive generative transformers," 2021, *arXiv:2105.14211v4*.
[75] Z. Wang, W. Liu, Q. He, X. Wu, and Z. Yi, "CLIP-GEN: Language-free training of a text-to-image generator with CLIP," 2022, *arXiv:2203.00386v1*.
[76] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," 2022, *arXiv:2205.11487v1*.
[77] Z. Feng, Z. Zhang, X. Yu, Y. Fang, L. Li, X. Chen, Y. Lu, J. Liu, W. Yin, S. Feng, Y. Sun, L. Chen, H. Tian, H. Wu, and H. Wang, "ERNIE-ViLG 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts," 2022, *arXiv:2210.15257v1*.
[78] Y. Balaji, S. Nah, X. Huang, A. Vahdat, J. Song, Q. Zhang, K. Kreis, M. Aittala, T. Aila, S. Laine, B. Catanzaro, T. Karras, and M.-Y. Liu, "EDiff-I: Text-to-image diffusion models with an ensemble of expert denoisers," 2022, *arXiv:2211.01324v3*.
[79] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10674–10685.

[80] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with CLIP latents," 2022, *arXiv:2204.06125*.

[81] J. Shi, C. Wu, J. Liang, X. Liu, and N. Duan, "DiVAE: Photorealistic images synthesis with denoising diffusion decoder," 2022, *arXiv:2206.00386v1*.

[82] W.-C. Fan, Y.-C. Chen, D. Chen, Y. Cheng, L. Yuan, and Y.-C. F. Wang, "Frido: Feature pyramid diffusion for complex scene image synthesis," 2022, *arXiv:2208.13753v1*.

[83] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation," 2022, *arXiv:2208.12242v1*.

[84] B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, and M. Irani, "Imagic: Text-based real image editing with diffusion models," 2022, *arXiv:2210.09276v1*.

[85] D. Valevski, M. Kalman, Y. Matias, and Y. Leviathan, "UniTune: Text-driven image editing by fine tuning an image generation model on a single image," 2022, *arXiv:2210.09477v3*.

[86] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo, W. Manassra, P. Dhariwal, C. Chu, Y. Jiao, and A. Ramesh. (2023). *Improving Image Generation With Better Captions*. [Online]. Available: https://cdn.openai.com/papers/dall-e-3.pdf

[87] W. Wu, Z. Li, Y. He, M. Zheng Shou, C. Shen, L. Cheng, Y. Li, T. Gao, D. Zhang, and Z. Wang, "Paragraph-to-image generation with information-enriched diffusion model," 2023, *arXiv:2311.14284*.

[88] W. Li, X. Xu, X. Xiao, J. Liu, H. Yang, G. Li, Z. Wang, Z. Feng, Q. She, Y. Lyu, and H. Wu, "UPainting: Unified text-to-image diffusion generation with cross-modal guidance," 2022, *arXiv:2210.16031v3*.

[89] R. Ganz and M. Elad, "CLIPAG: Towards generator-free text-to-image generation," 2023, *arXiv:2306.16805v2*.

[90] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, and Y. J. Lee, "GLIGEN: Open-set grounded text-to-image generation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jan. 2023, pp. 22511–22521.

[91] Y. Li, H. Wang, Q. Jin, J. Hu, P. Chemerys, Y. Fu, Y. Wang, S. Tulyakov, and J. Ren, "SnapFusion: Text-to-image diffusion model on mobile devices within two seconds," 2023, *arXiv:2306.00980v2*.

[92] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," 2023, *arXiv:2302.05543v2*.

[93] W. Zhao, Y. Rao, Z. Liu, B. Liu, J. Zhou, and J. Lu, "Unleashing text-to-image diffusion models for visual perception," 2023, *arXiv:2303.02153v1*.

[94] J. Hu, X. Han, X. Yi, Y. Chen, W. Li, Z. Liu, and M. Sun, "Efficient cross-lingual transfer for Chinese stable diffusion with images as pivots," 2023, *arXiv:2305.11540v1*.

[95] F. Ye, G. Liu, X. Wu, and L. Wu, "AltDiffusion: A multilingual text-to-image diffusion model," 2023, *arXiv:2308.09991v2*.

[96] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," 2017, *arXiv:1706.08500*.

[97] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2016, pp. 2234–2242.

[98] A. Borji, "Pros and cons of GAN evaluation measures," *Comput. Vis. Image Understand.*, vol. 179, pp. 41–65, Feb. 2018.

[99] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Mach. Learn. Res.*, vol. 139, 2021, pp. 8748–8763.

[100] C. Akkus, L. Chu, V. Djakovic, S. Jauch-Walser, P. Koch, G. Loss, C. Marquardt, M. Moldovan, N. Sauter, M. Schneider, R. Schulte, K. Urbanczyk, J. Goschenhofer, C. Heumann, R. Hvingelby, D. Schalk, and M. Aßenmacher, "Multimodal deep learning," 2023, *arXiv:2301.04856v1*.

[101] P. Wang. (2022). *Dall-E 2—PyTorch*. Accessed: Oct. 25, 2023. [Online]. Available: https://github.com/lucidrains/DALLE2-pytorch

SARAH K. ALHABEEB received the B.Sc. degree in information technology from the Department of Information Technology, College of Computer, Qassim University, Saudi Arabia, in May 2018, where she is currently pursuing the M.Sc. degree in information technology. Her research interests include machine learning, artificial intelligence, natural language processing, and the Internet of Things.

AMAL A. AL-SHARGABI received the master's and Ph.D. degrees from Universiti Teknologi Mara (UiTM), Malaysia. She is currently an Associate Professor with the College of Computer, Qassim University. She has been receiving a number of Qassim University's research grants, since 2018. Her research interests include program comprehension, empirical software engineering, and machine learning.