International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Advancements and Challenges in Text-to-Image Synthesis: A Comprehensive Review

Khushboo Patel[1] and Parth Shah[2]

Submitted: 25/01/2024 **Revised**: 03/03/2024 **Accepted**: 11/03/2024

Abstract: Text-to-image synthesis, a subfield of generative adversarial networks (GANs), is an exciting area of research that aims to bridge the gap between natural language understanding and computer vision. With recent advancements in deep learning techniques and the availability of large-scale datasets, significant progress has been made in generating realistic and diverse images from textual descriptions. Generating Hi-Fidelity, complex images from text are a challenging task. The ability to generate real images from textual descriptions has profound implications in various domains, including computer vision, multimedia, and virtual reality. This paper provides an in-depth study of state-of-the-art techniques and methodologies for text-to-image synthesis. Also, this paper discusses the various architectural enhancements, models, and evaluation metrics. Finally, the paper concludes by identifying open research issues and future directions that can enhance the performance and capabilities of text-to-image synthesis systems.

Keywords: *Text-to-image, GAN, Computer Vision,* Virtual Reality

1. Introduction

1. Text-to-image synthesis is a technology that aims to generate realistic images from textual descriptions. It involves using artificial intelligence and deep learning models to translate written text into visual representations. The goal is to create images that match the content and details described in the input text. The process of text-to-image synthesis typically involves two main tasks [32]: (i) understanding the text content i.e., NLP models are used to understand and interpret the textual descriptions. These models can extract relevant information, infer context, and comprehend the semantics of the text. (ii) representing the contextual form of the text and visualizing it in image form i.e., CV models are employed to generate images based on the extracted information from the NLP component. These models try to create visually coherent and plausible images that align with the given textual input. A typical diagram of text to image synthesis process is depicted in Fig 1.

Over the last decade, GAN models have witnessed the success in various applications including text-to-image synthesis, also known as generating visual content from textual descriptions. This domain lies at the intersection of computer vision and natural language processing, aiming to bridge the gap between language understanding and visual perception. Owing to the convenience of language for users, text-to-images synthesis has attracted many researchers and become an active research area. This promising field has wide-ranging applications, including content creation, virtual reality, augmented reality, and more. Advances in deep learning techniques, integrating with the availability of large-scale datasets, boosted the progress of text to image synthesis. Traditional approaches are relying on template-based methods or handcrafted rules to generate images from text. However, these methods often produced simplistic and limited results, lacking visual fidelity and semantic consistency. In recent years, generative models, particularly generative adversarial networks (GANs) and variational autoencoders (VAEs) have revolutionized text to-image synthesis. GANs, with their adversarial training framework, have shown remarkable progress in generating high-quality and visually appealing images. VAEs, on the other hand, excel at learning latent representations that capture the underlying semantics of textual descriptions. Furthermore, the amalgamation of GANs and VAEs has led to novel architectures that leverage the power of both models. The table-1 represents the past contribution in this field along the models involved. The review encompasses both GAN-based and VAE-based methods and their hybrid variants, discussing their ability to capture fine-grained object details, maintain spatial coherence, and ensure semantic consistency.

The key contributions of this paper are as follows:

1) We present a comprehensive review and analysis of text to image synthesis.

   We review existing text to image synthesis models and techniques that have been proposed in the

[1]*U and P U. Patel Department of Computer Engineering, Chandubhai S. Patel Institute of Technology (CSPIT), Faculty of Technology and Engineering (FTE), Charotar University of Science & Technology (CHARUSAT), Changa, Gujarat, India Email: khushboo30990@gmail.com,*

[2]*Smt. Kundanben Dinsha Patel Department of Information* Technology, *Charotar University of Science & Technology (CHARUSAT), Changa, Gujarat, India Email: parthshah.ce@charusat.ac.in*

literature, including architectural enhancements and various benchmarking.

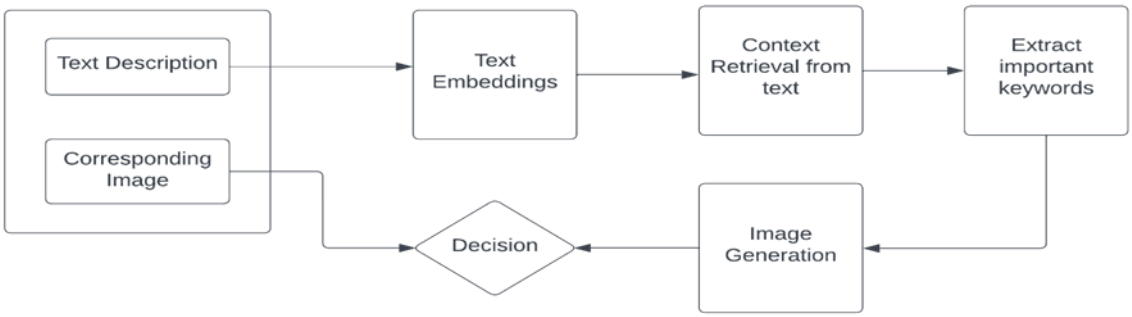

Figure 1. Typical architecture of a text to image synthesis system

2) We analyse the strengths and weaknesses of these techniques and highlight their applicability in different domains.

The rest of the paper is organized as follows. Section-II presents a detailed literature review based on taxonomy depicted in Fig. 1. Section-III discusses the evaluation metrics for text-to-image synthesis. Common datasets for this task are described in Section IV. The challenges and limitations are listed in Section-V, and applications of this task are provided in Section-VI. Tools and libraries required for implementing image synthesis from text are given in Section-VII. Finally, Section-VIII concludes the paper with future challenges.

2. Approaches for Text to Image Synthesis

This section provides an overview of the related work, highlighting the key advancements and challenges encountered in text-to-image synthesis. We first provide a taxonomy of text to image synthesis based on various models, architectural elements, evaluation metrics, and benchmarks. As the field progressed, researchers turned to generative models, particularly generative adversarial networks (GANs), to tackle the complexities of text-to-image synthesis. The Fig. 2 shows the taxonomy of the related work on which we discuss the existing literature.

As proposed in [32], text-to-image synthesis task can be addressed using three approaches:

1) General text-based image synthesis: In this kind of systems, the text description contains objects and the relationships between them.

2) Dialog text-based image synthesis: An interactive approach is used through multiple iterations of text descriptions in this method for generating images.

3) Scene layout-based image synthesis: In contrast to the first two approaches in which the quality of the generated image is relatively poor, this method aims to generate images with improved quality by adding conditional information to control the process.

A. Models

1) Generative Models

Reed et al. [1] proposed the Generative Adversarial Text-to-Image Synthesis (GAN-INT-CLS) model that utilized a GAN framework to generate images conditioned on text embeddings. The discriminator involved in GAN distinguishes between authentic and synthesized images and determine if the text and image were correctly paired. Inspired by the success of GANs, Zhang et al. [2] proposed the StackGAN framework that aimed to generate images hierarchically. The model consisted of a text-embedding conditioned stage-I GAN, which generated images at a coarse scale, followed by a stage-II GAN that refined the images at a finer scale. The results demonstrated the significant improvements in image quality and semantic consistency.

Xu et al. [3] proposed the AttnGAN model that incorporates the class conditioning. The author enhanced the fine-grained details in the synthesized images by extending the GAN-INT-CLS model, which makes use of an attention mechanism to align textual and visual data at various levels. This research made clear how crucial it is for text-to image synthesis to take both the global and local environment into account. Some advancements through different variants in the AttnGAN model have also been proposed in literature [9,29]. Text-to-image synthesis has also been investigated using variational autoencoders (VAEs). The VAE-GAN framework was suggested by Mansimov et al. [33], which integrated the advantages of both models. A continuous latent space was easier to learn thanks to the VAE component, enabling more fluid interpolations between textual descriptions. The GAN component's adversarial loss substantially enhanced the image quality and variety.

Fine-grained object features, spatial coherence, and semantic consistency are some of the issues that text-to

image synthesis researchers have recently been tackling in their study. The Mirror GAN model, which used a cutting edge two-stage generation approach, was first introduced by Xu et al. [8]. A basic picture layout was created in the first step, which was refined in the second stage using attention processes. The spatial coherence and fine-grained features in the synthesized images were enhanced by this method. While significant progress has been made in text-to-image synthesis, challenges remain. One such challenge is the generation of diverse and high-quality images. However, certain works have attempted to address generation of diversified images to some extent [5,18,22,24]. Achieving a balance between visual fidelity and semantic consistency remains a topic of ongoing research. Methods incorporating more precise control over the image generation process, such as conditional generation and attribute manipulation, are being explored to address this challenge. Another important aspect is the evaluation of synthesized images.

TABLE I. Related work on text to image synthesis

Author/Year	Model/Framework	Key Contributions
Reed et al. (2016)[1]	T2I GAN	Introduced the use of GANs for textto-image synthesis.
Zhang et al. (2017)[2]	StackGAN	Proposed a two-stage GAN architecture for generating high-resolution images with fine-grained details.
Xu et al. (2018)[3]	AttnGAN	Introduced attention mechanism to selectively attend to textual and visual cues, improving quality and diversity of synthesized images.
Razavi et al. (2019)[24]	VQ-VAE-2	Introduced vector quantization in VAEs for generating diverse and high-quality images conditioned on text.
Chen et al. (2020)[23]	DM-GAN	Employed dual-modal learning to jointly model text and image information, improving image synthesis quality.
Li et al. (2021)[22]	TediGAN	Introduced a text-driven intermediate network to bridge the semantic gap between text and images, improving synthesis quality and diversity.
Zhang et al. -2022	CoVAEGAN	Proposed a novel co-attention mechanism to enhance the alignment between text and visual features, producing more contextually relevant images.
Huang et al. (2023)[7]	ControlGAN	Introduced a controllable text-to-image synthesis model that allows users to manipulate visual attributes and generate images with specific characteristics.
Park et al. (2023)[19]	Context-aware GAN	Integrated textual context to guide image synthesis, improving semantic consistency and contextual relevance.

Commonly used metrics include image quality assessment and diversity analysis. However, these metrics may only capture part of the qualities required for meaningful evaluation. Developing comprehensive evaluation metrics that consider visual appeal and semantic consistency is an active area of research.

StackGAN [2] introduced a two-stage GAN architecture to address the challenge of semantic consistency and fine-grained details. The stage-I GAN generates a low-resolution image from text, which the stage-II GAN then refines to produce high-resolution and photo-realistic images. StackGAN++ [14] further improved the model's performance by incorporating a novel attention mechanism, allowing the generator to attend to specific textual cues and generate more contextually relevant images.

2) VAE based Models

In addition to GAN-based approaches, variational autoencoders (VAEs) have also been explored in text-to image synthesis. Text-to-Image Variational Autoencoder (T2I VAE) is a VAE-based model that decodes textual descriptions into related images by mapping them to a continuous latent space. With the help of the vector quantization technique provided by VQ-VAE-2 [24], VAEs may now generate a variety of high-quality, text-conditioned images. Additionally, hybrid models that combine the advantages of VAEs and GANs have grown in popularity. When creating an image, AttnGAN [3] presented a novel attention mechanism that allows users to selectively pay to various textual and visual stimuli. The quality and variety of synthetic images were greatly improved using AttnGAN by utilizing discriminative and generative learning. Text-to

image synthesis model evaluation is still difficult. To gauge the calibre and variety of synthetic images, conventional measures like Inception Score [34] and Frechet Inception´ Distance [35] have been modified. These measures, however, fall short in terms of semantic consistency and fine-grained details. In order to give a more thorough evaluation of synthesized pictures, current work have concentrated on generating innovative evaluation measures, such as R-precision [3] and Spatial Similarity Index (SSIM). Text-to image synthesis has come a long way, but there are still many obstacles to overcome. For instance, more research is needed to achieve controllability and diversity in synthesis, generate images with fine-grained details, and retain spatial coherence. Recent studies have also looked into how textual context and multi-modal learning might be combined to improve image synthesis. The table II presents the various models involved in text to image synthesis.

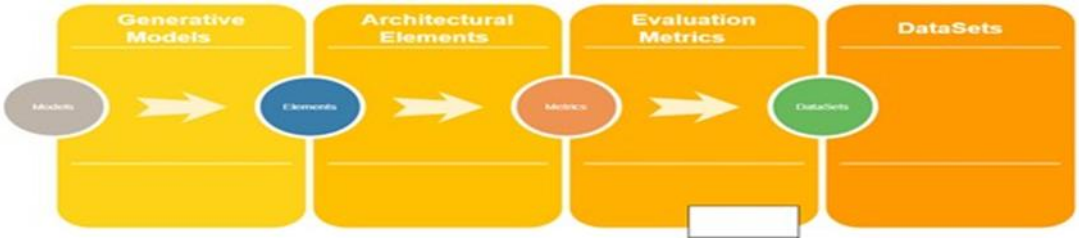

Figure 2. Taxonomy of the related work

B. Architectural Advancements

1) Attention Mechanism

AttnGAN, which includes an attention mechanism to specifically contribute to various textual and visual signals during the image synthesis process, was introduced by Xu et al. (2018) in [3]. This method significantly boosted the variety and calibre of synthetic images. CoVAEGAN, a co-attention mechanism that improved the alignment of text and visual characteristics, was proposed by Zhang et al. in 2022. The model produces more contextually relevant images by attending to specific textual cues and visual regions. Wang et al. introduced cross-modal VAE, a VAE-based model that incorporates a cross-modal attention mechanism to capture the semantic correspondence between text and image modalities. The model generates images by jointly modelling the textual and visual information. Self-Attention Generative Adversarial Network (SAGAN), proposed in [6] provided functionality of long-range dependency information use for generating images. In contrast to conventional GANs, the SAGAN could synthesize the image using characteristics from all feature locations.

2) Multimodal Fusion

Zhu et al. [4] employed dual-modal learning in DMGAN to jointly model text and image information. The model integrated a text-to-image synthesis branch with a text-to-attribute mapping branch to improve the quality and relevance of synthesized images. Zhang et al. [2] proposed the StackGAN architecture, a two-stage GAN model for generating high-resolution images from text. The Stage-I GAN generates a low-resolution image, which is then refined by the Stage-II GAN to produce high-quality images with fine-grained details.

3) Semantic Conditioning

Reed et al. [1] in 2016 proposed T2I VAE, a VAE based approach for text-to-image synthesis that maps textual descriptions to a continuous latent space. The model leverages word embeddings to capture semantic information. Reed et al. also introduced GAN-INT-CLS, a GAN-based model that incorporates attribute embeddings to bridge the semantic gap between text and images. The model improves the quality and diversity of synthesized images.

4)Controllable Synthesis

There have been some works [7,10,13,20,36] where researchers have addressed the concerns pertaining to the user control on the process of image generation so as to make it customized. Li et al. in [35] introduced ControlGAN, a controllable text-to-image synthesis model that allows users to manipulate visual attributes and generate images with specific characteristics. The model enables precise control over the generated images by conditioning on attribute vectors. AttnGAN [3] introduced conditional image generation by conditioning the model on specific textual descriptions. The attention mechanism selectively attends to the text embeddings, leading to the generation of contextually relevant images. Context-aware VAE leverages textual context to guide image synthesis. The model employs a hierarchical VAE framework that utilizes both global and local context information to generate images with improved semantic consistency.

3.Evaluation Metrics

For image quality assessment, a commonly used metric, e.g., Inception Score (IS), is often either mis calibrated for the single-object case or misused for the multi-object case.

TABLE II. GAN/VAE Models for Text to Image synthesis			
Author	**Approach**	**Key Contributions**	**Limitations**
Reed et al. [1]	GAN-based (T2I GAN)	Introduces GAN-based text to-image synthesis	Limited fine-grained details and visual realism
Zhang et al. [2]	GAN-based (StackGAN)	Two-stage GAN architecture for generating high resolution images	Lacks semantic consistency and context-awareness
Zhang et al. [14]	GAN-based (StackGAN++)	Incorporates attention mechanism for improved image generation	Limited diversity in synthesized images
Reed et al. [1]	VAE-based (T2I VAE)	Maps textual descriptions to continuous latent space for image generation	Challenges in maintaining semantic consistency
Razavi et al. [24]	VAE-based (VQ-VAE-2)	Introduces vector quantization technique for diverse image synthesis	Limited control over generated images
Salimans et al. [33]	Evaluation Metrics (Inception Score)	Measures the quality and diversity of synthesized images	Limited in capturing semantic consistency
Heusel et al. [34]	Evaluation Metrics (Frechet´ Inception Distance)	Evaluates the quality and diversity of generated images	Insensitive to fine-grained details

The current R-precision (RP) and Semantic Object Accuracy (SOA) metrics exhibit an overfitting phenomenon for text relevance and object accuracy assessment, respectively. Numerous crucial evaluation parameters, such as object integrity, positional alignment, and counting alignment, are generally disregarded in the case of multiple objects.

A. Traditional Metrics

1) Inception Score(IS)

Salimans et al. [34] in 2016 introduced the Inception Score, which measures the quality and diversity of generated images based on their class probabilities predicted by an Inception model. Higher Inception Scores indicate better quality and diversity.

2) Fre´chet Inception Distance (FID)

Heusel et al. [35] in 2017 proposed the Frechet Inception´ Distance, which computes the distance between the feature representations of real and generated images using an Inception model. Greater similarity and quality are indicated by lower FID values.

3) Multi-Scale Structural Similarity (MS-SSIM)

Wang et al. in 2004 used SSIM to assess the structural similarity between real and generated images. It measures perceptual quality by considering luminance, contrast, and structural information. This is a method of measuring image similarity based on the measurement of luminance, contrast and structure. The structural similarity (SSIM) calculation between images x and y is shown in Eq. 3-A3.

$$SS\ IM(x,y) = [l(x,y)^{\alpha} * c(x,y)^{\beta} * s(x,y)^{\gamma}] \quad (1)$$

Where, $l(x,y)^{\alpha}$, $c(x,y)^{\beta}$, and $s(x,y)^{\gamma}$ are luminance factor, contrast factor and structure factor, respectively. The MSSSIM uses scaled images for the similarity evaluation to incorporate the images with different resolutions. Higher value of MS-SSIM indicates better performance.

4) R-Precision

Reed et al. [1] in 2016 introduced R-precision, which measures the precision of retrieved images for a given query text. It quantifies how well the generated images match the intended textual descriptions.

5) Visual Semantic Similarity

This measure attempts to fill the gap between semantic consistency evaluation between the text and synthetic image. This subjective metric is computed as the feature vector distance between the synthetic image and the textual description.

4. Datasets

The datasets and various resources like pretrained models form a basis for text-to-image synthesis research and facilitate experimentation, comparison, and progress in text to-image synthesis. The following are the common available dataset from various resources depicted in Table III.

5. Open Issues and Challenges

In this section, we discuss various issues and challenges involved in text to image synthesis as described in Fig. III

TABLE III. Various resources and dataset for text-to-image synthesis

Sr. No.	Dataset/Resources	Description
1	COCO (Common Objects in Context) [25]	Contains images with diverse object categories includes rich annotations and captions. benchmark for text-to-image synthesis.
2	Oxford-102 Flowers [26]	Contains102 categories of flowers, often used for fine-grained text-to-image synthesis tasks focuses on flower species.
3	CUB - 200 - 2011 (Caltech - UCSD Birds - 200 - 2011	It contains 11,788 images of 200 subcategories belonging to birds, 5,994 for training and 5,794 for testing.
4	Multi-ModalCelebA-HQ	Following CelebA-HQ, 30,000 high-resolution face images from the CelebA collection were selected. High-quality segmentation masks, sketches, descriptive text, and translucent backgrounds are on each image.
5	LAION-COCO	The largest dataset of 600M high-quality subtitles for public web photos.
6	MS COCO Captions	An extension of the COCO dataset which provides textual descriptions corresponding to each image.
7	Conceptual Captions [28]	Contains over three million images with diverse concepts and descriptive captions.

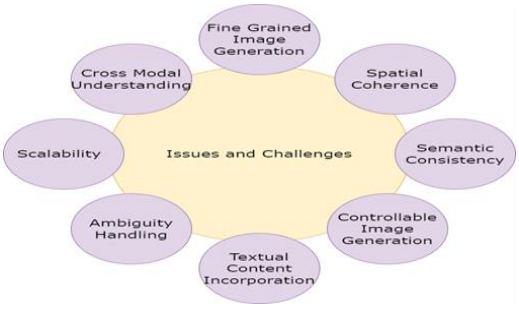

Figure 3. Issues and Challenges

A. Fine-Grained Details

Generating images with fine-grained details, such as textures, intricate patterns, and small objects, remains a challenge in text-to-image synthesis. Current models often need help to capture and reproduce such information accurately.

B. Consistency and Coherence

Ensuring consistency and coherence between the textual descriptions and generated images is an ongoing challenge. It involves accurately translating the semantics of the text into visual elements and maintaining coherent relationships between objects [16,17].

C. Handling Ambiguity

Textual descriptions sometimes leave possibility for more than one possible interpretation. Text-to-image synthesis models ought to take this uncertainty into account and provide visuals that accurately represent the intended meaning or, as necessary, offer multiple interpretations.

Figure 4. Applications of text-to-image synthesis

D. Scalability

It is difficult to scale up text-to-image synthesis to handle enormous datasets and produce high-resolution images. As a result, effective techniques that may produce high quality photographs within a fair amount of time and with limited resources are required.

6. Applications of Text-to-Image Synthesis

This section is focused on applications and use cases of text to image generation in various sectors. Fig. 4 shows some crucial fields where this task is prominently in demand.

In the design and advertising sectors, text-to-image synthesis can be used to swiftly create visual material for goods, logos, adverts, and branding. It makes it possible for designers to effectively transfer ideas from text into visual representations. By automatically creating visual scenes or pictures based on textual descriptions, it can help with storytelling and illustration. It might therefore be used in children's novels, comics, and online storytelling platforms.

A. Virtual Environments and Gaming

It is possible to use text-to-image synthesis to develop realism and immersion in VR and AR experiences. Based on textual descriptions, it enables the creation of visually appealing locations, items, and characters. It can help game creators create a variety of realistic gaming elements, including people, places, and things. It improves the visual quality and variety of gaming worlds and enables effective content generation.

B. E-commerce and Product Visualization

By creating visual representations of the product concepts given in the text, text-to-image synthesis can help in product design and prototyping. Before physical prototyping, it enables quick visualization and iteration of designs. By creating realistic product images from text descriptions, it can improve the online purchasing experience. As a result, it enhances the overall user experience and aids customers in visualizing things before making judgments about purchases.

C. Generative Art

Generative art can be produced using text-to-image synthesis, in which literary descriptions act as inspiration for the creation of one-of-a-kind, emotionally charged visual works of art. It combines language and visuals in fresh, creative ways. It gives artists and designers a platform to transform their literary conceptions and ideas into visual forms. It opens up new avenues for artistic expression and increases the potential for creative experimentation.

D. Data Augmentation and Image Generation

Object recognition, image captioning, and scene comprehension are just a few of the computer vision tasks that can benefit from text-to-image synthesis. It enables additional training data with diverse visual content and annotations.

7. Tools and Libraries

The following are the various tools and libraries that helps to develop applications based to text to image synthesizing.

TensorFlow: An open-source machine learning framework that provides a flexible platform for building and training deep learning models. TensorFlow offers various modules and APIs that are commonly used in text-to-image synthesis research.

PyTorch: Another popular deep learning framework that provides a dynamic computational graph and efficient GPU acceleration. PyTorch offers many tools and modules for building text-to-image synthesis models and conducting research experiments.

GANs (Generative Adversarial Networks) Libraries: Several GANs libraries are available that provide preimplemented architectures and components for training GAN models. Examples include:

PyTorch-GAN: A PyTorch-based library that implements various GAN variants, including those used in textto-image synthesis. TensorFlow-GAN: A TensorFlow library specifically designed for training GAN models, including those used in text-to-image synthesis research.

NLTK (Natural Language Toolkit): A Python package that offers resources and methods for handling data from human language. For text preparation, tokenization, partof-speech tagging, and other natural language processing activities involved in text-to-image synthesis, NLTK provides a variety of features.

OpenAI GPT: Text-to-image synthesis research has used OpenAI's Generative Pre-trained Transformer (GPT) models, such as GPT-2 and GPT-3, as prompts for text generation and image synthesis. The models can be used with the Hugging Face Transformers library or using the OpenAI API.

Hugging Face Transformers: An extensive collection of pre-trained transformer models, including GPT and other language models, are available in a Python package. It provides simple-to-use APIs for text generating and can be used in research involving text-to-image synthesis.

TorchVision: A PyTorch library that offers tools for computer vision, such as datasets, models, and transformations, that are often used. Text-to-image synthesis models can be utilized with TorchVision's pre-trained models and tools for image processing.

PIL (Python Imaging Library) or Pillow: Python libraries for image processing and manipulation. These libraries provide features for loading, resizing, cropping, and other picture modifications that are frequently needed in research involving text-to-image synthesis.

Scikit-image: A Python library for computer vision and image processing operations. Research on text-to-image synthesis can benefit from a variety of methods offered by Scikit-image for image editing, filtering, segmentation, and other operations.

These tools and libraries facilitate developing, experimenting, and evaluating text-to-image synthesis models. In addition, researchers often utilize combinations of these tools based on their preferences and the specific requirements of their research projects.

8. Conclusion and Future Directions

To create photorealistic images that are semantically compatible with the text descriptions, text-to-image synthesis (T2I) is used. We provide a taxonomy that groups the material that has already been published according to a number of criteria, such as model architectures, generative adversarial networks (GANs) models, variational autoencoder (VAE) models, architectural upgrades, evaluation measures, and benchmarking methods. We then provided detailed discussions on each taxonomy category, summarizing the key contributions and innovations in the respective subfields. Finally, we explore the advancements in GAN models for text-to-image synthesis, including techniques such as conditional GANs, attention mechanisms, and progressive training. Additionally, we discuss the dataset and resource available and finally we have showcased various issues and challenges involved in recent research. Finally, the future work advocates the design of models that can address the challenges of image quality, semantic understanding, and controllability, and by exploring multimodal and cross domain synthesis, the field can advance towards generating high-quality, contextually relevant, and user-centric visual content from textual descriptions.

References

[1] Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H. (2016, June). Generative adversarial text to image synthesis. In International conference on machine learning (pp. 1060-1069). PMLR.

[2] Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D. N. (2017). StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).

[3] Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., Metaxas, D. N. (2018). AttnGAN: Fine-Grained Text to Image

Generation with Attentional Generative Adversarial Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[4] Zhu, J. Y., Zhang, R., Zhang, D., Lu, J., Ziwei, L., Luo, X. (2019). DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-to-Image Synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[5] Shen, T., Zhou, T., Long, G., Jiang, J., Zhang, C. (2019). Diverse Image Generation via Self-conditioned GANs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[6] Zhang, H., Goodfellow, I., Metaxas, D., Odena, A. (2019). SelfAttention Generative Adversarial Networks. In Proceedings of the International Conference on Machine Learning (ICML).

[7] Li, Y., Liu, S., Yang, J., Zhou, X. (2019). Controllable Text-to Image Generation. In Proceedings of the AAAI Conference on

Artificial Intelligence (AAAI).

[8] Chen, Y., Yang, Z., Yang, Y., Zhang, M., Zhang, J. (2020). Mirror GAN: Learning Text-to-image Generation by Redescription. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[9] Wang, T., Wang, M., Liu, J., Zhu, J. Y., Tao, A., Kautz, J., Catanzaro, B. (2020). AttnGAN++: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. In Proceedings of the European Conference on Computer Vision (ECCV).

[10] Liu, S., Zhu, Z., Li, N., Luo, X., Shi, J. (2021). CPVT: A Compact Progressive Text-to-Image Synthesis Model with Fine-Grained User Control. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[11] Chen, T., Zhang, Z., Zhang, J. (2019). MirrorGAN: Learning textto-image synthesis by redescription. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 4091-4100).

[12] Zhu, J. Y., Park, T., Isola, P., Efros, A. A. (2017). Unpaired imageto-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 2223-2232).

[13] Tan, H., Chan, C. S., Agustsson, E., Veeling, B. S. (2019). TextGAN++: A consistent and controlled text-to-image generative adversarial network. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, pp. 9594-9601).

[14] Zhang, Y., Zhang, Z., Xu, J., Zhang, Z. (2019). StackGAN++: Realistic image synthesis with stacked generative adversarial networks. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 41(8), 1947-1962.

[15] Huang, X., Li, Y., Poursaeed, O., Hopcroft, J., Belongie, S. (2018). Multimodal unsupervised image-to-image translation. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 179-196).

[16] Chen, T., Zhang, Z., Zhang, J. (2020). Semantics disentangling for text-to-image generation. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 752-769).

[17] Wang, T., Zhu, M., Torr, P. H. (2020). Towards high-resolution text-to-image synthesis with pixel-wise semantic alignment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 7297-7306).

[18] Dai, B., Zhang, L., Wang, D. (2017). Towards diverse and natural image descriptions via a conditional GAN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 2970-2979).

[19] Wu, Z., Li, Z., Fan, Z. G., Wu, Y., Gan, Y., Pu, J., Li, X. (2023). Learning Monocular Depth in Dynamic Environment via Context aware Temporal Attention. arXiv preprint arXiv:2305.07397.

[20] Tao, M., Bao, B. K., Tang, H., Xu, C. (2023). GALIP: Generative Adversarial CLIPs for Text-to-Image Synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14214-14223).

[21] Zhou, H., Qi, L., Huang, H., Yang, X., Wan, Z., Wen, X. (2022). CANet: Co-attention network for RGB-D semantic segmentation. Pattern Recognition, 124, 108468.

[22] Xia, W., Yang, Y., Xue, J. H., Wu, B. (2021). Tedigan: Text-guided diverse face image generation and manipulation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2256-2265).

[23] Zhu, M., Pan, P., Chen, W., Yang, Y. (2019). Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5802-5810).

[24] Razavi, A., Van den Oord, A., Vinyals, O. (2019). Generating diverse high-fidelity images with vq-vae-2. Advances in neural information processing systems, 32.

[25] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13 (pp. 740-755). Springer International Publishing.

[26] Nilsback, M. E., Zisserman, A. (2008, December). Automated flower classification over a large number of classes. In 2008 Sixth Indian Conference on Computer Vision, Graphics Image Processing (pp. 722-729). IEEE.

[27] Fang, H., Gupta, S., Iandola, F., Srivastava, R. K., Deng, L., Dollar,´ P., ... Zweig, G. (2015). From captions to visual concepts and back. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1473-1482).

[28] Sharma, P., Ding, N., Goodman, S., Soricut, R. (2018, July). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 2556-2565).

[29] Naveen, S., Kiran, M. S. R., Indupriya, M., Manikanta, T. V., Sudeep, P. V. (2021). Transformer models for

enhancing AttnGAN based text to image generation. Image and Vision Computing, 115, 104284.

[30] Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D. N. (2017). Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In Proceedings of the IEEE international conference on computer vision (pp. 5907-5915).

[31] Kliger, M., Fleishman, S. (2018). Novelty detection with gan. arXiv preprint arXiv:1802.10560.

[32] Zhou, R., Jiang, C., Xu, Q. (2021). A survey on generative adversarial network-based text-to-image synthesis. Neurocomputing,451, 316-336.

[33] Mansimov, Elman, et al." Generating images from captions with attention." arXiv preprint arXiv:1511.02793 (2015).

[34] Salimans, Tim, et al." Improved techniques for training gans." Advances in neural information processing systems 29 (2016).

[35] Heusel, Martin, et al." Gans trained by a two time-scale update rule converge to a local nash equilibrium." Advances in neural information processing systems 30 (2017).

[36] Li, Bowen, et al." Controllable text-to-image generation." Advances in Neural Information Processing Systems 32 (2019).

Khushboo Patel is a research scholar at Charusat university, Changa, Gujarat. Her research area includes machine learning and deep learning including generative models. She has published book chapter on international level.

Dr Parth Shah received his PhD in cloud computing at the Charusat University, Changa, Gujarat. He is the head and professor at the Department of Information Technology, Charusat University, Changa, Gujarat. His research interests include issues related to machine learning and deep learning, cloud computing, Internet of things and cyber security. He is author of many journal papers, conference papers and book chapters which are published at international and national level.