International Conference on Recent Trends in Data Science and its Applications
DOI: rp-9788770040723.129

Synthetic Image Generation from Text Data

Saksham Bansal
Department of Data Science and Business Systems SRM Institute of Science and Technology,
KattankulathurChennai, Tamil Nadu,
India saksham002bansal@gmail.com

S. Sharanya
Assistant Professor,Department of Data Science and Business Systems SRM Institute of Science and Technology,
KattankulathurChennai-603203,
Tamil Nadu, India sharanys1@srmist.edu.in

DeekshaRawat
Department of Data Science and Business Systems SRM Institute of Science and Technology,
KattankulathurChennai, Tamil Nadu,
India deeksharawat0002@gmail.com

***Abstract*—Automatic synthesis of realistic images is challenging, and even state-of-the-art artificial intelligence and machine learning algorithms suffer from not fulfilling this expectation. However, the emergence of image processing has allowed operations on an image to enhance or extract information from it and synthesize pictures from textual descriptions, which has become an active research area in recent times. The already-developed model by OpenAI surprised the world after its launch. However, everything worthwhile has a price. Further study is necessary since the model could not account for problems including gender prejudice, stereotypes, language structure, viewpoint, writing, symbolism, and the delivery of explicit material. This survey report aims to supplement past studies using different image processing techniques to create synthetic images. This article critically assesses current approaches to assess text-to-image synthesis models, draws attention to the existing architectures' limitations, and identifies new research areas. To further advance research in the field, improvement of the architectural design and model training is needed. This can be achieved by developing better datasets and evaluation metrics.**

***Index Terms*—Synthetic image generation, Text-to-image, GANs, DALL-E, Imagen, Image processing**

I. Introduction

Logic will get us from point A to B, whereas imagination will get us everywhere. Imagination is the soil that brings a dream to life, the preview of life's coming attraction. The mind starts imagining scenarios as soon as a voice is heard or a text is read. For example, "woof woof," without effort; the mind presented a picture of a dog. This is what the mind is capable of, thinking about the unknown. Sometimes things can get uncanny or unrealistic. For example, "teddy bears grocery shopping in ancient times," seen in Figure 1. This scenario cannot come to life. However, with the advent of technology, realistic images of such unrealistic thoughts can be produced.

Computer vision has made a significant breakthrough in giving rise to what the world looks like. Whether autonomous vehicles, facial recognition, medical imaging, manufacturing, education, or even transportation, the impact has been significant for each domain. It allowed to get meaningful information from the surroundings, whether the input is through images,videos, or even a live feed through a camera, and then perform the required tasks and take the appropriate action.

Fig. 1. "Teddy bears shopping for groceries in ancient Egypt" developed using DALL-E 2 [4].

Computer vision enables computers to recognize objects in videos and images the same way humans do. The advancement in domains like deep learning, artificial intelligence and innovations such as neural networks have enabled computers to transcend the capability of humans. Better computing power is needed to process the generated data with each passing day, as the daily generated data can reach as high as 2.4 quintillion bytes (that is seventeen zeroes). Since the start, objects have been classified, but the accuracy associated with datasets needs improvement from time to time.

Digital image processing, which allows operations to be performed on pictures by converting them to digital form, is one of the latest advancements in the field of computer vision. Recently, image generation has taken over the world and is the most highly researched domain of image processing. Im- age generation involves generating new images from already existing datasets.

The emergence of Generative Adversarial Networks (GANs)[2] provided an approach for generative models to further advance the image generation domain using deep learning methods. The model tries to learn and discover patterns in input data and generate an output that combines images from the pre-existing dataset. Some examples of other generative models could be Na¨ıve Bayes, Deep Belief Network, LatentDirichlet Allocation, Variationa Autoencoder, Gaussian Mix- ture Model, Restricted Boltzmann Machine, and the Gaussian Mixture Model. A representation of the GAN architecture may be seen in the Figure 2.

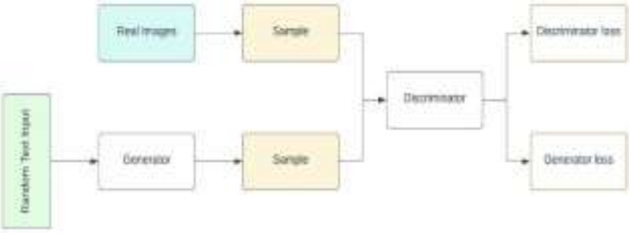

Fig. 2. GAN Architecture.

The GANs were further developed with the conditional operation to generate the required output. The condition is provided through class values or labels. The discriminator is then provided with the conditional input and the picture (genuine or false).

This survey aims to highlight the evolution of text-to-image conversion models across time. The paper highlights the research done and tries to showcase the advantages and disadvantages of the various methodologies and datasets that have been used.

II. Components

This section revisits three critical components needed to un- derstand the text-to-image approaches in the following areas: GAN [2] for synthesising images from their text descriptions, text encoders are used to produce the embedded text for training to map the prompt to a representation space and datasets commonly used by the text-to-image community.

A. Generative Adversarial Network

Two Neural Networks, a Generator network and a Discrim- inator network, made up the basic GAN [2] models. The discriminator is trained to discriminate between actual and fraudulent pictures produced by the model during training. The generator is trained to collect the genuine data distribution and create pictures in order to trick the discriminator. A representation of the GAN architecture may be seen in the Figure 3.

More technically, as shown in [2], it is a min-max optimization formulation in which the Generator wants to reduce the objective function. Simultaneously, the Discriminator seeks to maximise the same objective function. The Discriminator wishes to reduce the probability of D(G(z)) to zero. As a result, it seeks to maximize (1-D(G(z))). In contrast, the Generator seeks to push the probability of D(G(z)) to 1, causing the Discriminator to incorrectly identify the created sample as genuine. As a result, the Generator wishes to minimize (1- D(G(z)). Instead of providing merely noise as input to Gen- erator, it turns the textual description into a text embedding, concatenate it with a noise vector, and then provide it as input to Generator.

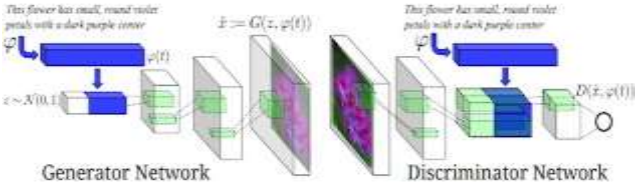
This flower has small, round violet petals with a dark purple center
φ
$\varphi(t)$
$z \sim \mathcal{N}(0,1)$
$\hat{x} := G(z, \varphi(t))$
This flower has small, round violet petals with a dark purple center
φ
$D(\hat{x}, \varphi(t))$
Generator Network
Discriminator Network

Fig. 3. Text-conditional convolutional GAN architecture [20].

B. Textual Embedding

Producing an embedded text from textual expressions that the model may use as a training variable is crucial. Recently, automatically trained generalizable and highly discriminative text representations from words and characters have been developed using deep convolutional and recurrent networks for text [20]. A textual description is encoded in [19] utilising a hybrid character-level convolutional recurrent neural network that has been trained (char-CNN-RNN). These works have proved to be an inspiration to discover a mapping that goes straight from words and characters to visual pixels.

The authors of StackGAN [32] suggested Conditioning Augmentation (CA) that uses a ”hybrid character-level convolutional recurrent neural network,” the same as [20]. ”GAN Text to Image Synthesis” paper. Rather than using the fixed text embedding obtained by a pre-trained text encoder, it randomly picks latent variables from an independent Gaus- sian distribution, where the matrix of covariance and mean are functions of the text embedding. This method produces additional training pairs and encourages smoothness in the latent conditioning manifold. This technology was used by many of the text-to-image techniques that followed.

Instead of the char-CNN-RNN, which extracts semantic vectors from text descriptions, the inventors of AttnGAN[30] used a bi-directional Long Short-Term Memory (LSTM) [26]. In the bidirectional LSTM, each word corresponds to two hidden states, one in each direction. As a result, the two hidden states are combined to reflect a word’s semantic meaning. The development of an attention mechanism for the generator allows it to draw various subregions of the picture by concentrating on phrases that are most relevant to the sub-region being drawn. Precisely, each term in the sentence is encoded into a vector representation. On the other hand, the linguistic depiction is stored as a global sentence vector. Furthermore, a deep attentional multimodal similarity model is illustrated to compute image-text matching loss at a granular level.

C. Datasets

Every machine-learning challenge is built on datasets. CUB- 200-2011 Birds [3], Oxford-120 Flowers [14], and COCO[10] are commonly used datasets in text-to-image research. Caltech-UCSD Birds-200-2011 (CUB-200-2011) is an ex- panded version of the CUB-200 dataset, with about twice as many photos per class and additional component location annotations. Oxford-102 Flower is a dataset with 102 flowerclassifications. Each image displays a single item and is ac- companied by ten captions. COCO [10] dataset is a significant resource for analysing object identification, classification, and interpretation with around 120k pictures and five captions per image. Images from the COCO dataset usually depict several regularly interacting objects in complex contexts, making the environment more complicated than the Oxford-102 Flowers and CUB-200-2011 Birds datasets. Table I summarizes the dataset statistics. Additional datasets used were the Multi- Modal-CelebA-HQ dataset [29], the CelebA-Dialog dataset [7], the FFHQ-Text dataset [35], and the CelebAText-HQ dataset [27]. The majority of text-to-image works employ the official 2014 COCO split.

III. Methods

Following the last chapter’s discussion of GANs, text en- coders, widely used datasets. State-of-the-art approaches for direct text-to-image creation are discussed further. The first text-to-image technique was presented in 2016 by Reed et al. [20], followed by Stack Generative Adversarial Networks [32]. The introduction of AttnGAN [30], the diffusion model [11], and the usage of CLIP architecture [15] will be discussed next.

A. First text to image approaches

Reed et al [20] established the first text-to-image technique, which produced a simple and successful model for creating images based on precise visual descriptions. The description embedding is compressed to a tiny dimension using a fully- connected layer, followed by leaky-ReLU, and then concate- nated to the noise vector. The method involves conditioning a deep convolutional generative adversarial network (DC- GAN) on text characteristics encoded by a hybrid character- level RNN. Feed-forward learning was conducted by both the generator and discriminator networks, followed by batch normalization [37] on all convolutional layers.

GAN-CLS was presented in [20] to handle the multi-modality issue in text-to-image generation, which combines improvements in DCGAN with an RNN encoder to create pic- tures from a latent variable and embedding image descriptions. On the other hand, GAN-CLS [20] fails to produce credible representations of more complex and variable realistic scenar- ios, such as those depicting human activities. A representation of the GAN-CLS architecture may be seen in the Figure 4.

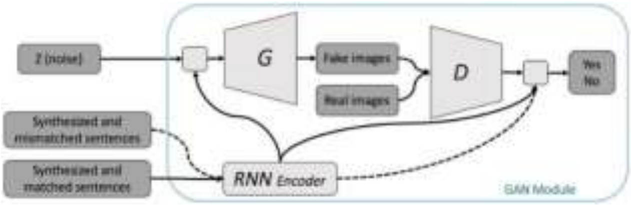
Z (noise)
G
Fake images
Real images
D
Yes
No
Synthesized and mismatched sentences
Synthesized and matched sentences
RNN Encoder
GAN Module

Fig. 4. GAN-CLS for text-to-image synthesis [36].

Instead of binary discrimination, GAN-CLS [20] trains the discriminator to distinguish between three conditions: actual pictures with matched text, false images with random text, and real images with mismatched text.

B. *StackGAN*

Early models [20] could generate only 64×64 images based on the text description provided, which led to a lack of detail and clarity in the produced images, which further led to the development of stack generators to synthesize higher- resolution images. The authors of the StackGAN [32] proposed the two stages image generation method. The first stage, Stage- I GAN, generates low-resolution images by sketching elementary shapes and colors associated with the text description provided. The second stage, Stage-II GAN, takes the results produced in the first stage along with the text description given and produces photo-realistic images with a higher resolution. The images generated are 256×256, much higher than those generated by previous models. The Stage-I GAN needs to stabilize conditional GAN [2] training to improve the generated samples' diversity and add randomness to the network. For this purpose, it uses Conditioning Augmentation (CA). It makes the generator network robust by capturing intricate details of the object achieved by introducing more image-text pairs. A representation of the StackGAN architecture may be seen inthe Figure 5.

The proposal of StackGAN [32] or StackGAN-v1 [32] came with a demand for conditional and unconditional task generators in the form of StackGAN++ [33] or StackGAN-v2 [33]. It involved sharing of parameters between multiple generators arranged in a tree-like structure. This improved the clarity of generated images even further and made the training stable. Color consistency regularization was then introduced to expedite multi-distribution approximation. Overall, StackGAN[32] proved to outperform the previous methods that generated photo-realistic images.

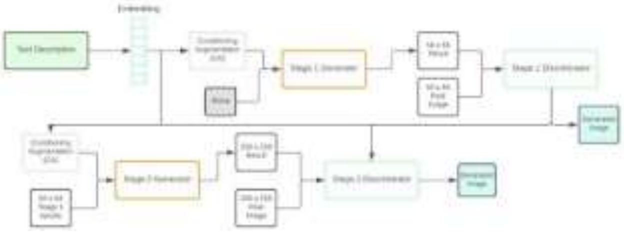

Fig. 5. StackGAN Architecture.

C. *AttnGAN*

The recent text-to-image generative models are based on Generative Adversarial Networks (GANs) [2]. However, the essential information is lost from the input sentences, and the images produced are of low resolution. The generation of high- quality images and detailed oriented analysis of the provided text was needed. For this purpose, Attentional Generative Adversarial Network (AttnGAN) [30] was developed as shown in Figure 6.

Deep Attentional Multimodal Similarity Model (DAMSM) and an Attentional Generative Network are the two maincomponents of the model. The former component focuses on the most relevant words of the text description by developing an attention mechanism for the generator by drawing different sub-regions. An attention layer forms a word-context vector using an image vector to query the word vectors present in each sub-region. This forms a multimodal context vector that generates new image features in surrounding sub-regions, thus giving an image with more details and higher resolution. The latter component computes similarities between the sentence and generated image. It produces an attention-driven image- text matching score using two encoders, a text encoder and an image encoder.

TABLE IMAJOR DATASETS USED FOR TEXT-TO-IMAGE SYNTHESIS

Dataset	Training Images	Testing Images	Total Images	Captions per Image	Object Categories
COCO	82783	40504	123287	5	80
CUB-200 Birds	8855	2933	11788	10	200
Oxford-102 Flowers	7034	1155	8189	10	102

With AttnGAN's original architecture, control over the location of objects and the identity of objects could not be obtained. To achieve this, an object pathway (OP) was added to the AttnGAN [30]. This AttnGAN+OP [6] made it easier to get desired objects at desired locations.

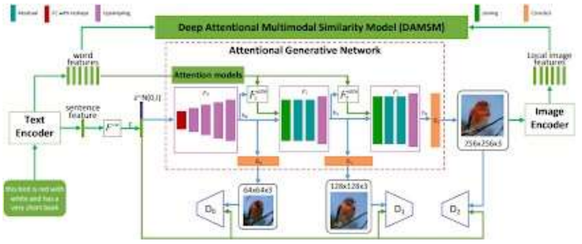

Fig. 6. AttnGAN Architecture [30].

D. Diffusion Model

The idea behind diffusion models is to create a noisy image by adding a little amount of Gaussian noise to a photograph. Then it repeats the procedure, adding more Gaussian noise to the picture to get an even noisier image. This process is repeated numerous times (up to 1000 times) to get a noisy picture. Then train a neural network using the noisier sample as input and the job of predicting the denoised version of the picture as output. The Diffusion architecture is shown in Figure 7.

The author of the VQ-Diffusion model [11] suggested an approach based on a vector-quantized variational autoencoder that uses a conditional variant of the Denoising Diffusion Probabilistic Mode (DDPM) to find hidden space is the best way to reduce noise in an image. This method is well-suited for text-to-image generation tasks as it eliminates the single- direction bias in current techniques and integrates an iffusion strategy to reduce error, which is a significant issue with existing systems. As a result, the VQ-Diffusion model [11]outperforms conventional auto-regressive (AR) [31] models in text-to-image creation.

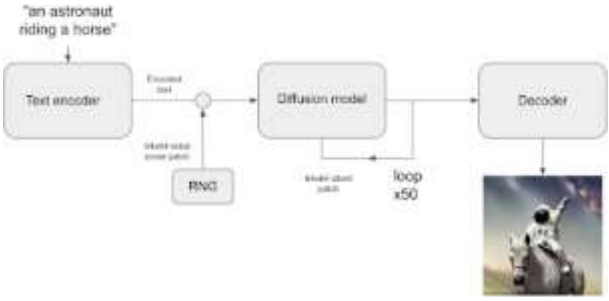

Fig. 7. Illustration of Diffusion model [16].

Dreambooth [21] describes a unique approach for customis- ing text-to-image diffusion models (specialising them to user requests). Given a few images of a topic, it fine-tunes a pre-trained text-to-image model to tie a unique identification code with the particular subject. The unique identification may create whole new realistic visual of the subject contextualised in varying circumstances after the particular subject is included into the model's output domain. This approach synthesizes the subject in varied settings, positions and viewpoints that are not clear in the sampling pictures by using the lexical stored in the model and a new uniaxial class-specific retention loss.

E. CLIP Architecture

Scaling models on massive datasets of annotated pictures acquired from the internet have been a driving force behind recent development in the field of computer vision. Within the confines of this paradigm, CLIP [15] has established itself as an adequate representation learner for pictures. CLIP [15] embeddings offer several desired qualities, including being resistant to image distribution shifts, having excellent zero- shot capabilities, and is fine-tuned to deliver state-of-the-art outcomes on a broad range of visual tasks. It trains a text encoder and an image encoder to predict the right couplings of a collection of (text, image) training samples. It first calculates the image's feature embedding and the collection of potential texts' feature embedding via their respective encoders. The cosine similarity of these embeddings is then determined, adjusted by a temperature parameter, and normalized by a softmax into a probability distribution. The text representation with the highest similarity score will be selected as the best representation of the image's content. Refer to Figure8. Optimizing the latent space of a GAN [2] can produce pictures with high semantic relevance to the input text [34].Compared to standard benchmarks, methods combining GAN[2] and CLIP [15] are training-free and zero-shot,requiring no expensive or specialised training data. However, the image space of the CLIP+GAN [1] technique is restricted by the pre- trained GAN [2]. This makes generating images with unusual object pairings, which are not present in the GAN's training data, challenging.

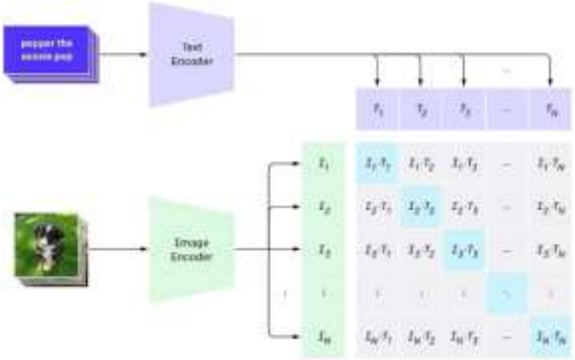

Fig. 8. CLIP pre-training to predict picture-text pairs [15].

The authors of fusedream [12] proposed a composite generation approach that enhances two pictures such that they may be seamlessly combined to make a natural and contextually suitable images. It effectively uses a novel dynamic barrier gradient descent approach to transform a composite generation into a special bi-level optimisation problem that boosts the AugCLIP [12] score while including an interoceptive consistency score as a secondary aim.

IV. Discussion

In the previous chapters, cutting-edge text-to-image meth- ods, commonly utilized techniques, and architectures were looked at. Following that, the current progress in this subject will be outlined and the existing problems will be identified. Image synthesis from text has significantly developed in contrast to a straightforward design in 2016 [20], which used a basic GAN loss during training, consisting of a generator and discriminator. Models based on various methodologies were presented and trained on massive datasets of text-image pairings. Some techniques, however, depend on pre-trained models, such as Generative Adversarial Networks, which search across the generative model's latent space using a gradient-based strategy to update the latent vector, relying on loss functions such as cosine similarity. Modern approaches often include a multi-stage pipeline and numerous contributing losses.

A. Limitations

As part of the continuing preview of this technology, summarize initial findings on potential risks associated with existing approaches and measures that aim to alleviate these concerns. Disseminating these findings promotes a greater understanding of image creation, alteration technologies, andassociated risks. Without adequate safeguards, the image- generating models might be used to create a wide variety of false or otherwise damaging content, which could influence how people perceive the veracity of information in general. Some existing models inherit different biases from their training data, and their outputs might occasionally reinforce social stereotypes.

Although the generated image is consistent with the description as a whole, individual image regions or parts of some things are frequently not recognizable or compatible with the words in the sentence, such as ”a white crown.” This major limitation is revealed when one examines the generated images in greater detail. The author of SSA-GAN [9] proposed a new framework called Semantic-Spatial Aware GAN to generate pictures from input text to solve this challenge.

Many current algorithms no longer provide any result on the Oxford-102 Flowers dataset. Evaluating Text to image algorithms on a single object dataset using CUB-200-2011 Birds [3] is sufficient, while Oxford-102 Flowers [14] did not offer more valuable insights. Another possibility is to utilize the CelebA-HQ dataset [27] for text-to-image approaches.

The evaluation of created pictures' level of excellence, diversity, and linguistic alignment is a challenging and continuous problem. The emergence of IS [25] and FID [5] has made it simpler, although they have flaws. Aside from the IS and FID, several other approaches have been made, including the classification accuracy score (CAS) [18], the density and coverage metrics [13], the detection-based score [22], precision and recall metrics [8] [24], and SceneFID [28].

V. Conclusion And Future Work

This survey presented a synopsis of the various methodologies and advancements in synthetic image generation from text data. The various datasets that have been used over the years and the changes that have arrived with each was discussed. Novel structures and techniques have been presented and tested on real-world data using massive media datasets. The proposed solutions were compared and the challenges that are yet to be resolved were discussed.

The early models proposed GAN [2] as the solution for text- to-image conversion, but it did not always produce the desired results. With the advent of diffusion models, the generator's aim is to fool the discriminator and reverse the image from noise. Although it needed much computational power, it paved the way for developing advanced models, skyrocketing the amount of fame for image processing.

Future developments in this domain involve the input being an image and a text, video, or speech. Recent research has emphasized the conversion of speech to image, text to video, and text to a 3-dimensional object or shape. We offer a novel inpainting technique for obtaining particular semantic characteristics regarding corrupted regions by contrasting sections with complementary picture and informative text. It lets you modify specific areas of an image by displaying a mask and a text prompt specifying what to replace. To increase the semantic closeness of the produced picture and the text,an image-text matching loss is used. This study will assist researchers to better comprehend the current state-of-the-art field and the unresolved difficulties that remain.

References

[1] K. Crowson, S. Biderman, D. Kornis, D. Stander, E. Hallahan, L. Castri- cato, and E. Raff, “Vqgan-clip: Open domain image generation and edit- ing with natural language guidance,” arXiv preprint arXiv:2204.08583, 2022.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D.W. Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” Communications of the ACM, vol. 63, pp. 139–144, 2020.

[3] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona, “Caltech-UCSD Birds 200”. California Institute of Technology. CNS-TR-2010-001. 2010.

[4] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, “Hierarchical text-conditional image generation with clip latents,” arXiv preprint arXiv:2204.06125, 2022.

[5] 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,” Advances in neural information processing systems, vol. 30, 2017.

[6] T. Hinz, S. Heinrich, and S. Wermter, “Generating multiple objects at spatially distinct locations,” arXiv preprint arXiv:1901.00686, 2019.

[7] Y. Jiang, Z. Huang, X. Pan, C.C. Loy, and Z. Liu, “Talk-to-edit: Fine-grained facial editing via dialog,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 13799–13808.

[8] T. Kynkaanniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila, “Improved precision and recall metric for assessing generative models,” Advances in Neural Information Processing Systems, vol. 32, 2019.

[9] W. Liao, K. Hu, M. Y. Yang, and B. Rosenhahn, “Text to image generation with semantic-spatial aware GAN,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 18187–18196.

[10] T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in European conference on computer vision, Springer, 2014, pp. 740– 755.

[11] 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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10696–10706.

[12] X. Liu, C. Gong, L. Wu, S. Zhang, H. Su, and Q. Liu, ”Fusedream: Training-free text-to-image generation with improved clip+ gan space optimization,” arXiv preprint arXiv:2112.01573, 2021.

[13] M. F. Naeem, S. J. Oh, Y. Uh, Y. Choi, and J. Yoo, “Reliable fidelity and diversity metrics for generative models,” in International Conference on Machine Learning, 2020, pp. 7176–7185.

[14] M.-E. Nilsback, and A. Zisserman, “Automated flower classification over a large number of classes,” in 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, IEEE, 2008, pp. 722– 729.

[15] 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 supervi- sion,” in International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.

[16] I.Stenbit, F. Chollet, and L. Wood, “A walk through latent space with stable diffusion.” keras.io. https://keras.io/examples/generative/random walkswith stable diffusion/ (accessed October 15, 2022).

[17] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, “Zero-shot text-to-image generation,” in

International Conference on Machine Learning, PMLR, 8821–8831, p. 2021.

[18] S. Ravuri, and O. Vinyals, "Classification accuracy score for conditional generative models," Advances in neural information processing systems, vol. 32, 2019.

[19] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 49–58.

[20] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Gen- erative adversarial text to image synthesis," in International conference on machine learning, PMLR, 2016, pp. 1060–1069.

[21] 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," arXiv preprint arXiv:2208.12242, 2022.

[22] S. Sah, D. Peri, A. Shringi, C. Zhang, M. Dominguez, A. Savakis, and

R. Ptucha, "Semantically invariant text-to-image generation," in 2018 25th IEEE International Conference on Image Processing (ICIP), IEEE, 2018, pp. 3783–3787.

[23] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. SeyedGhasemipour, 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," arXivpreprint arXiv:2205.11487, 2022.

[24] M. S. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, and S. Gelly, "Assessing generative models via precision and recall," Advances in neural information processing systems, vol. 31, 2018.

[25] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and

X. Chen, "Improved techniques for training gans," Advances in neural information processing systems, vol. 29, 2016.

[26] M. Schuster, and K. K. Paliwal, "Bidirectional recurrent neural networks," IEEE transactions on Signal Processing, vol. 45, pp. 2673–2681, 1997.

[27] J. Sun, Q. Li, W. Wang, J. Zhao, and Z. Sun, "Multi-caption text-to-face synthesis: Dataset and algorithm," in Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 2290–2298.

[28] Sitharthan, R., Vimal, S., Verma, A., Karthikeyan, M., Dhanabalan, S. S., Prabaharan, N., ...&Eswaran, T. (2023). Smart microgrid with the internet of things for adequate energy management and analysis.Computers and Electrical Engineering, 106, 108556.

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

[30] 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 Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1316–1324.

[31] J. Yu, Y. Xu, J. Y. 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," arXiv preprint arXiv:2206.10789, 2022.

[32] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D.

N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 5907–5915.

[33] 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 transactions on pattern analysis and machine intelligence, vol. 41, pp. 1947–1962, 2018.

[34] Z. Zhang, and L. Schomaker, "OptGAN: Optimizing and Interpreting the Latent Space of the Conditional Text-to-Image GANs," arXiv preprint arXiv:2202.12929, 2022.

[35] Y. Zhou, and N. Shimada, "Generative Adversarial Network for Text-to- Face Synthesis and Manipulation with Pretrained BERT Model," in 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition FG 2021, IEEE, 2021, pp. 01–08.

[36] H. Dong, J. Zhang, D. McIlwraith, and Y. Guo, "Learning text to image synthesis with textual data augmentation," arXiv preprint arXiv:1703.06676, p. 2, 2017.

[37] Moshika, A., Thirumaran, M., Natarajan, B., Andal, K., Sambasivam, G., &Manoharan, R. (2021).Vulnerability assessment in heterogeneous web environment using probabilistic arithmetic automata. IEEE Access, 9, 74659-74673.